

La freccia del tempo cognitivo

Irreversibilità, entropia e il costo termodinamico del cambiamento

Quinta monografia del ciclo *Architettura della mente*

Nota di chiusura sul ciclo

Questa monografia completa il ciclo *Architettura della mente* rispondendo alla domanda che le quattro precedenti hanno sistematicamente eluso pur incontrandola a ogni svolta: perché la cognizione ha una freccia del tempo?

Il ciclo ha costruito un'architettura precisa. La prima monografia ha mostrato che il malinteso è il punto fuori equilibrio del gioco comunicativo — ma non ha domandato perché recuperare la comprensione condivisa dopo un malinteso costi più che stabilirla la prima volta. La seconda ha formalizzato la co-evoluzione substrato-linguaggio come sistema dinamico accoppiato con attrattori stabili — ma non ha domandato perché il consolidamento della memoria sia asimmetrico rispetto alla sua formazione. La terza ha costruito la torre bayesiana gerarchica e i limiti dell'auto-conoscenza formale — ma non ha domandato perché aggiungere un livello alla torre costi meno che toglierlo. La quarta ha analizzato il semigruppo dell'attenzione e la viscosità dell'identità cognitiva — ma non ha domandato perché la viscosità sia sempre positiva, perché nessun sistema cognitivo sia reversibile.

La risposta a tutte queste domande è la stessa: la cognizione è un processo fisico che avviene in un sistema termodinamico fuori dall'equilibrio, e la freccia del tempo cognitiva è la freccia del tempo termodinamica applicata all'aggiornamento delle credenze. La produzione di entropia è il prezzo fisico di ogni aggiornamento, il costo di Landauer è il pavimento termodinamico di ogni atto di dimenticare, il teorema di fluttuazione di Crooks quantifica l'asimmetria tra apprendere e disapprendere, e la disuguaglianza di Jarzynski connette il lavoro cognitivo gratuito con la variazione di energia libera nella distribuzione delle credenze.

L'apparato formale viene dalla termodinamica stocastica — Seifert (2012), Crooks (1999), Jarzynski (1997), Sagawa e Ueda (2010), Parrondo, Horowitz e Sagawa (2015) — e dalla geometria dell'informazione di non-equilibrio. È un apparato distinto da quello di ciascuna delle quattro monografie precedenti eppure strutturalmente connesso con tutte: la divergenza KL ricompare come funzione di Lyapunov per la produzione di entropia, la metrica di Fisher ricompare come tensore di diffusione per le fluttuazioni di non-equilibrio, e lo schema diagonale di Lawvere ricompare nel

dèmoni di Maxwell cognitivi come struttura dell'impossibilità dell'auto-raffreddamento epistemico.

Proemio: il tempo che non torna

Un fatto banale e quasi mai formalizzato: imparare qualcosa è diverso dal disimpararlo. Non diverso nel senso vago che la psicologia del dimenticare differisce da quella dell'apprendere — diverso nel senso preciso che i due processi hanno direzioni termodinamiche opposte rispetto all'equilibrio. Imparare porta il sistema fuori dall'equilibrio in una direzione; dimenticare non è semplicemente il percorso inverso, non è il film girato al contrario. È un processo che produce a sua volta entropia, che costa lavoro, che lascia tracce irreversibili nella struttura del substrato.

Questa asimmetria è ovunque nell'esperienza cognitiva. Il trauma non si disimpara come si disimpara una regola grammaticale. La conoscenza procedurale dell'esperto non si perde come si perde un numero di telefono. L'identità cognitiva consolidata in anni di esperienza non si modifica con la stessa facilità con cui è stata costruita. La comprensione profonda di un argomento resiste all'oblio meglio della memorizzazione superficiale. In tutti questi casi c'è qualcosa che va oltre la semplice differenza di forza del ricordo — c'è una struttura asimmetrica nel tempo, una direzione preferita che la formalizzazione termodinamica rende precisa.

La termodinamica classica affronta l'irreversibilità attraverso la seconda legge: in un sistema isolato, l'entropia non diminuisce. Ma i sistemi cognitivi non sono isolati — sono sistemi aperti che scambiano continuamente informazione e lavoro con l'ambiente. Per questi sistemi, la termodinamica classica è insufficiente: si ha bisogno della termodinamica stocastica, che tratta sistemi fuori dall'equilibrio soggetti a fluttuazioni stocastiche e forzati da protocolli variabili nel tempo. La termodinamica stocastica non afferma che l'entropia di un sistema aperto non possa diminuire localmente — afferma che la produzione totale di entropia, sistema più ambiente, non può essere negativa. Questa formulazione è quella corretta per i sistemi cognitivi: un agente può ridurre localmente la propria incertezza attraverso l'aggiornamento bayesiano, ma lo fa dissipando entropia nell'ambiente attraverso il processo fisico che realizza quell'aggiornamento.

L'argomento si sviluppa in sei movimenti. Il primo costruisce il ponte formale tra l'aggiornamento bayesiano e la termodinamica stocastica attraverso la connessione Jarzynski-Bayes — l'identità che mostra come la regola di Bayes sia un caso particolare del teorema di fluttuazione di Jarzynski. Il secondo analizza la produzione di entropia nell'equazione di Fokker-Planck per le credenze, mostrando che la divergenza KL è una funzione di Lyapunov per la produzione di entropia cognitiva. Il terzo applica il teorema di fluttuazione di Crooks alle traiettorie di aggiornamento per formalizzare l'asimmetria tra apprendimento e disapprendimento. Il quarto affronta il principio di Landauer applicato alla cancellazione delle credenze e il suo legame con il costo termodinamico del dimenticare e della riconsolidazione. Il quinto tratta il dèmoni di

Maxwell cognitivo — la domanda se la metacognizione possa "raffreddare" le credenze riducendo l'entropia cognitiva senza pagarne il costo — e mostra che non può, per ragioni strutturalmente connesse ai teoremi di impossibilità della terza monografia. Il sesto chiude il ciclo analizzando gli stati stazionari di non-equilibrio come struttura del trauma, dell'identità rigida e della polarizzazione collettiva.

Capitolo I — Il ponte Jarzynski-Bayes: la regola di Bayes come teorema di fluttuazione

I.1 La termodinamica stocastica: premesse necessarie

La termodinamica stocastica studia sistemi fuori dall'equilibrio soggetti a forze stocastiche, descrivendo la loro dinamica attraverso processi di Markov in tempo continuo o mappe stocastiche in tempo discreto. Il protagonista formale è il *lavoro stocastico* W — la quantità di lavoro compiuto sul sistema lungo una singola traiettoria stocastica — che fluttua da traiettoria a traiettoria e la cui distribuzione porta informazione sull'irreversibilità del processo.

Per un sistema con Hamiltoniana $H(x; \lambda)$ che dipende da un parametro di controllo λ variato secondo un protocollo $\lambda(t)$ da λ_A a λ_B , il lavoro compiuto lungo la traiettoria $x(t)$ è:

$$W[x(\cdot)] = \int_0^T (\partial H / \partial \lambda)(x(t); \lambda(t)) \cdot (d\lambda/dt) dt$$

La variazione di energia libera tra gli stati di equilibrio corrispondenti ai valori estremi del parametro è:

$$\Delta F = F(\lambda_B) - F(\lambda_A) = -\beta^{-1} \log (Z(\lambda_B)/Z(\lambda_A))$$

dove $Z(\lambda) = \int \exp(-\beta H(x; \lambda)) dx$ è la funzione di partizione e $\beta = 1/(k_B T)$ è il fattore inverso di temperatura. La seconda legge nella sua forma classica afferma $\langle W \rangle \geq \Delta F$: il lavoro medio non è inferiore alla variazione di energia libera. L'uguaglianza vale solo per processi quasi-statici — processi reversibili all'equilibrio in ogni istante.

I.2 Il teorema di Jarzynski: lavoro gratuito e fluttuazioni

Il teorema di Jarzynski (1997) è il risultato centrale della termodinamica stocastica fuori dall'equilibrio. Afferma che, indipendentemente dalla velocità del protocollo e quindi dall'ammontare di irreversibilità:

$$\langle \exp(-\beta W) \rangle = \exp(-\beta \Delta F)$$

La media dell'esponentiale del lavoro, presa su tutte le traiettorie stocastiche del processo, è l'esponentiale della variazione di energia libera. Questa è un'identità esatta — non un'approssimazione valida vicino all'equilibrio — che vale per qualsiasi processo stocastico markoviano con condizioni iniziali di equilibrio.

La dimostrazione standard usa la reversibilità microscopica (detailed balance) della dinamica di equilibrio. Sia $P[x(\cdot)]$ la probabilità della traiettoria diretta $x(t)$ e $\tilde{P}[\tilde{x}(\cdot)]$ quella della traiettoria coniugata temporale $\tilde{x}(t) = x(T-t)$ sotto il protocollo invertito $\tilde{\lambda}(t) = \lambda(T-t)$. Il principio di detailed balance generalizzato afferma:

$$P[x(\cdot)] / \tilde{P}[\tilde{x}(\cdot)] = \exp(\beta(W[x(\cdot)] - \Delta F))$$

il rapporto tra la probabilità della traiettoria diretta e quella della traiettoria inversa è l'esponentiale della dissipazione irriducibile $W - \Delta F$. Integrando su tutte le traiettorie:

$$\int P[x(\cdot)] \exp(-\beta W[x(\cdot)]) \mathcal{D}x = \exp(-\beta \Delta F) \cdot \int \tilde{P}[\tilde{x}(\cdot)] \mathcal{D}\tilde{x} = \exp(-\beta \Delta F)$$

che è esattamente il teorema di Jarzynski. La disuguaglianza di Jensen $\langle e^X \rangle \geq e^{\langle X \rangle}$ applicata al teorema di Jarzynski recupera la seconda legge: $\langle W \rangle \geq \Delta F$.

I.3 La connessione Jarzynski-Bayes: la regola di Bayes come identità di non-equilibrio

La connessione tra il teorema di Jarzynski e la regola di Bayes è uno dei risultati più profondi della termodinamica dell'informazione. La struttura formale è la seguente.

Si considera un sistema con stato $h \in H$ — un'ipotesi — e un parametro di controllo λ che rappresenta il segnale osservato $o \in \Omega$. La distribuzione di equilibrio a $\lambda = \lambda_{\text{prior}}$ è il prior $\pi(h)$; la distribuzione di equilibrio a $\lambda = \lambda_o$ è la distribuzione aggiornata $\pi(h|o)$. La variazione di energia libera tra i due stati di equilibrio è:

$$\Delta F(o) = F(\lambda_o) - F(\lambda_{\text{prior}}) = -\beta^{-1} \log (Z(\lambda_o)/Z(\lambda_{\text{prior}}))$$

dove $Z(\lambda_{\text{prior}}) = \sum_h \exp(-\beta H(h; \lambda_{\text{prior}})) \propto 1$ (normalizzazione del prior) e $Z(\lambda_o) \propto P(o)$ — la verosimiglianza marginale dell'osservazione. Quindi:

$$\Delta F(o) = \beta^{-1} \log P(o)$$

Il lavoro compiuto lungo la traiettoria singola $h(t)$ — l'ipotesi "attiva" nel sistema cognitivo durante l'aggiornamento — è:

$$W[h(\cdot)] = \int_0^T (\partial H / \partial \lambda)(h(t); \lambda(t)) (d\lambda / dt) dt$$

Per un protocollo che porta λ da λ_{prior} a λ_o attraverso la funzione di Hamiltoniana $H(h; \lambda) = -\beta^{-1} \log P(o|h) \cdot f(\lambda)$ — dove f cresce da 0 a 1 come funzione del tempo — il lavoro lungo la traiettoria che rimane nell'ipotesi h per tutto il protocollo è:

$$W[h] = -\beta^{-1} \log P(o|h)$$

la log-verosimiglianza negativa dell'osservazione sotto l'ipotesi h . Il teorema di Jarzynski afferma allora:

$$\langle \exp(-\beta W) \rangle_{\pi} = \sum_h \pi(h) \cdot \exp(\beta \cdot \beta^{-1} \log P(o|h)) = \sum_h \pi(h) P(o|h) = P(o) = \exp(-\beta \Delta F(o))$$

e questa è esattamente l'identità di marginalizzazione che produce $P(o)$ dalla regola di Bayes. Il *teorema di Jarzynski applicato agli aggiornamenti cognitivi* afferma che la verosimiglianza marginale $P(o)$ è la media dell'esponenziale del lavoro cognitivo — la somma pesata per il prior delle esponenziali della log-verosimiglianza — e questo è formalmente identico all'identità di Jarzynski con $\Delta F = \beta^{-1} \log P(o)$.

La regola di Bayes è il caso deterministico zero-temperatura ($\beta \rightarrow \infty$) del teorema di Jarzynski applicato agli aggiornamenti cognitivi: l'aggiornamento bayesiano ideale è il processo nel quale tutta la "massa" dell'ipotesi è trasportata verso l'ipotesi con massima verosimiglianza in modo quasi-statico, senza dissipazione. Qualsiasi aggiornamento reale che si discosti dall'ottimo bayesiano — incluso l'aggiornamento attenzionato parziale della quarta monografia — produce dissipazione cognitiva: lavoro in eccesso rispetto alla variazione di energia libera.

I.4 Il lavoro cognitivo gratuito e la dissipazione

Dalla connessione Jarzynski-Bayes segue una decomposizione precisa del lavoro cognitivo in parte "utile" e parte "dissipata". Per un agente che aggiorna le credenze da π a $\pi(\cdot|o)$, la variazione di energia libera cognitiva è:

$$\Delta F_{\{\text{cog}\}}(o) = \beta^{-1} \log P(o) = \beta^{-1} \log \sum_h \pi(h) P(o|h)$$

Questa è la quantità minima di lavoro cognitivo necessaria per produrre l'aggiornamento ottimale: il "lavoro gratuito" che può essere estratto dall'osservazione o nel senso termodinamico. Il lavoro effettivamente compiuto nel processo di aggiornamento reale è sempre maggiore o uguale a $\Delta F_{\{\text{cog}\}}$:

$$\langle W_{\{\text{eff}\}} \rangle \geq \Delta F_{\{\text{cog}\}}(o)$$

La differenza $\langle W_{\{\text{eff}\}} \rangle - \Delta F_{\{\text{cog}\}}(o) = \beta^{-1} D_{\text{KL}}(\pi_{\{\text{eff}\}}(\cdot|o) \parallel \pi(\cdot|o))$ è la dissipazione cognitiva: l'energia extra spesa per produrre la distribuzione aggiornata

π_{eff} anziché quella bayesiana ottimale $\pi(\cdot | o)$. Per l'aggiornamento attenzionato con maschera A della quarta monografia, la dissipazione è:

$$\Sigma_{\text{dissp}}^A = \beta^{-1} D_{\text{KL}}(\pi^A(\cdot | o) \parallel \pi(\cdot | o))$$

il costo termodinamico dell'attenzione selettiva misurato in unità di lavoro cognitivo. La dissipazione è proporzionale alla divergenza KL tra la distribuzione prodotta dall'attenzione parziale e quella bayesiana ottimale — esattamente la perdita di calibrazione del paragrafo I.4 della quarta monografia, ora interpretata come dissipazione fisica irreversibile.

I.5 La struttura della dissipazione come misura dell'irrazionalità

Il legame tra dissipazione cognitiva e divergenza KL permette di interpretare le deviazioni dalla razionalità bayesiana come dissipazione termodinamica. Un agente bayesianamente ottimale ha $\Sigma_{\text{dissp}} = 0$: aggiorna le credenze senza dissipazione, trasportando la massa di probabilità esattamente dove la verosimiglianza indica con il minimo lavoro. Un agente con confirmation bias — maschera $A_t \propto \pi_t$ dalla quarta monografia — ha dissipazione:

$$\Sigma_{\text{dissp}}^{\text{confirm}} = \beta^{-1} D_{\text{KL}}(\pi^{\text{confirm}}(\cdot | o) \parallel \pi(\cdot | o)) > 0$$

quantificabile esplicitamente in funzione della struttura della maschera e del gradiente bayesiano. Questo non è un giudizio normativo sulla razionalità ma un fatto fisico: gli aggiornamenti sub-ottimali costano più lavoro fisico degli aggiornamenti ottimali, e questo costo si manifesta come calore dissipato nel substrato biologico che realizza l'aggiornamento — un surplus di attivazione neurale, una maggiore spesa metabolica, una struttura di firing meno efficiente.

Il legame quantitativo tra inefficienza bayesiana e costo metabolico è supportato da evidenze neurobiologiche: la corteccia prefrontale ventromediale, sede primaria dell'aggiornamento delle credenze, mostra un consumo metabolico proporzionale alla difficoltà dell'aggiornamento — dove "difficoltà" corrisponde precisamente alla distanza tra la distribuzione aggiornata e quella ottimale bayesiana, cioè alla dissipazione Σ_{dissp} . Il framework termodinamico offre la struttura formale per questo fenomeno empirico, connettendo la psicologia cognitiva con la neuroscienza computazionale attraverso la termodinamica stocastica.

Capitolo II — Produzione di entropia e l'equazione di Fokker-Planck cognitiva

II.1 La produzione di entropia come funzionale del flusso di credenze

Nell'equazione di Fokker-Planck per le credenze individuali della quarta monografia:

$$\partial_t \pi_t = -\nabla_P \cdot (\pi_t \cdot b_t) + \beta^{-1} \Delta_P \pi_t$$

il flusso di probabilità è $J_t = \pi_t \cdot b_t - \beta^{-1} \nabla_P \pi_t$, dove b_t è il drift deterministico. La produzione totale di entropia per unità di tempo è:

$$\dot{s}_{\text{tot}}(t) = \dot{s}_{\text{sys}}(t) + \dot{s}_{\text{env}}(t)$$

dove $\dot{s}_{\text{sys}}$ è il tasso di cambiamento dell'entropia del sistema cognitivo e $\dot{s}_{\text{env}}$ è il tasso di produzione di entropia nell'ambiente (il substrato fisico). Per la dinamica di Fokker-Planck, questi si scrivono esplicitamente:

$$\dot{s}_{\text{sys}}(t) = -d/dt \int_P \pi_t \log \pi_t d\pi = -d/dt H(\pi_t)$$

il negativo del tasso di cambiamento dell'entropia di Shannon della distribuzione delle credenze, e:

$$\dot{s}_{\text{env}}(t) = \beta \int_P \|J_t(\pi)\|_{G^{-1}}^2 / \pi_t d\pi \geq 0$$

l'integrale del flusso di probabilità pesato dall'inversa della matrice di diffusione — sempre non-negativo per la struttura della metrica di Fisher. Il tasso totale di produzione di entropia è:

$$\dot{s}_{\text{tot}}(t) = \dot{s}_{\text{sys}}(t) + \dot{s}_{\text{env}}(t) = \beta \int_P (J_t \cdot G^{-1} \cdot J_t) / \pi_t d\pi - d/dt H(\pi_t) \geq 0$$

per la disuguaglianza di Cauchy-Schwarz applicata al flusso J_t .

Il risultato cruciale è la *decomposizione dell'entropia totale* in termini di divergenza KL dalla distribuzione stazionaria π^* :

$$d/dt D_{\text{KL}}(\pi_t \| \pi^*) = -\dot{s}_{\text{tot}}(t) + \dot{s}_{\text{house}}(t)$$

dove $\dot{s}_{\text{house}}(t) \geq 0$ è il *housekeeping entropy production* — la produzione di entropia necessaria per mantenere lo stato stazionario lontano dall'equilibrio termodinamico. Per sistemi all'equilibrio termodinamico, $\dot{s}_{\text{house}} = 0$ e la divergenza KL è una funzione di Lyapunov pura: decresce monotonicamente fino a zero. Per sistemi fuori dall'equilibrio — inclusi tutti i sistemi cognitivi biologici — $\dot{s}_{\text{house}} > 0$ e la divergenza KL non è più una funzione di Lyapunov globale: il sistema può

stabilizzarsi in stati stazionari lontani dall'equilibrio che richiedono un flusso continuo di energia per essere mantenuti.

II.2 La decomposizione di Hatano-Sasa: excess e housekeeping

Hatano e Sasa (2001) hanno dimostrato una decomposizione fondamentale della produzione di entropia per sistemi stazionari di non-equilibrio che generalizza il risultato di Seifert. Per ogni traiettoria $x(t)$ di un sistema con stato stazionario non-equilibrio $q^{\{ss\}}(x)$, la produzione di entropia si decompone in:

$$\Sigma_{\text{tot}}[x(\cdot)] = \Sigma_{\text{ex}}[x(\cdot)] + \Sigma_{\text{hk}}[x(\cdot)]$$

dove la *excess entropy production* è:

$$\Sigma_{\text{ex}}[x(\cdot)] = \int_0^T [\partial_t \log q^{\{ss\}}(x(t); \lambda(t))] dt$$

e la *housekeeping entropy production* è:

$$\Sigma_{\text{hk}}[x(\cdot)] = \Sigma_{\text{tot}}[x(\cdot)] - \Sigma_{\text{ex}}[x(\cdot)]$$

La excess entropy production misura il costo di guidare il sistema da uno stato stazionario all'altro attraverso la variazione del protocollo $\lambda(t)$: è il costo della transizione tra stati stazionari di non-equilibrio. La housekeeping entropy production misura il costo di mantenere il sistema in uno stato stazionario lontano dall'equilibrio: è il costo permanente dello stato di non-equilibrio, il tributo continuo alla seconda legge per esistere in uno stato organizzato.

Applicata alla cognizione: ogni sistema cognitivo biologico ha un *housekeeping cognitive entropy production* $\dot{\Sigma}_{\text{hk,cog}} > 0$ che rappresenta il costo metabolico continuo del mantenimento della struttura organizzata delle credenze lontana dall'equilibrio termico. Questo è il costo basale del pensiero — non il costo di aggiornare le credenze ma il costo di avere credenze strutturate in primo luogo. La *excess cognitive entropy production* è il costo aggiuntivo di transitare da una struttura di credenze a un'altra: è la formalizzazione termodinamica del costo dell'apprendimento, della terapia, della conversione, della riqualificazione professionale.

La decomposizione di Hatano-Sasa connette il framework termodinamico con il semigrupphe dell'attenzione della quarta monografia. La misura invariante μ^* del semigrupphe è lo stato stazionario di non-equilibrio del sistema cognitivo: l'identità cognitiva stabile che richiede un flusso continuo di attenzione (lavoro cognitivo) per essere mantenuta. Cambiare identità — spostarsi da μ_1 a μ_2 attraverso una transizione attenzionale — produce la excess entropy production Σ_{ex} , che è la misura termodinamica del costo del cambiamento. La viscosità dell'identità $\zeta = 1/\lambda_1$ della quarta monografia è inversamente proporzionale alla facilità con cui Σ_{ex} può

essere prodotto: identità ad alta viscosità richiedono grandi perturbazioni — grandi produzioni di excess entropy — per essere cambiate.

II.3 La divergenza KL come funzione di Lyapunov per la produzione di entropia cognitiva

Si dimostra ora che la divergenza KL dalla distribuzione stazionaria è una funzione di Lyapunov per la produzione di entropia cognitiva in una forma più precisa di quanto la quarta monografia avesse stabilito.

Per l'equazione di Fokker-Planck cognitiva con stato stazionario π^* :

$$d/dt D_{KL}(\pi_t \parallel \pi^*) = -I_F(\pi_t) + I_{\{hk\}}(\pi_t)$$

dove $I_F(\pi_t)$ è la *Fisher information* di π_t relativa a π^* che misura la dissipazione verso l'equilibrio, e $I_{\{hk\}}(\pi_t)$ è il contributo del housekeeping. Esplicitamente:

$$I_F(\pi_t) = \int P \parallel \nabla_P \log(\pi_t/\pi^*) \parallel^2_{\{G(\pi_t)\}} \pi_t d\pi \geq 0$$

dove la norma è nella metrica di Fisher $G(\pi_t)$. Questa è l'informazione di Fisher nella forma usuale ma valutata sul rapporto di densità π_t/π^* anziché sulla densità stessa.

Il prodotto scalare della geometria di Fisher con il gradiente del log-rapporto ha un'interpretazione precisa: è la velocità con cui la distribuzione delle credenze si avvicina alla distribuzione stazionaria nella metrica dell'informazione. Per sistemi senza housekeeping (stati di equilibrio), $I_{hk} = 0$ e si ha la disuguaglianza:

$$d/dt D_{KL}(\pi_t \parallel \pi^*) \leq -I_F(\pi_t) \leq 0$$

che è esattamente la monotonia della divergenza KL verso lo stato di equilibrio — la funzione di Lyapunov della quarta monografia come caso particolare. Per sistemi con housekeeping, il termine $I_{\{hk\}}$ può essere positivo e produrre stati stazionari con $D_{KL}(\pi^{\{ss\}} \parallel \pi_{\{eq\}}) > 0$: il sistema è stabilizzato lontano dall'equilibrio con produzione continua di entropia. Questo è il fondamento termodinamico degli stati cognitivi persistenti — compreso il trauma, come si vedrà nel Capitolo VI.

II.4 Il tasso di produzione di entropia come metrica della vitalità cognitiva

La quantità $\dot{s}_{tot}(t)$ — il tasso totale di produzione di entropia cognitiva — è una misura della *vitalità cognitiva* del sistema: un sistema che produce alta entropia sta elaborando informazione attivamente, aggiornando le credenze, mantenendo strutture lontano dall'equilibrio. Un sistema cognitivo in declino — verso la rigidità cognitiva, la demenza, la stasi — mostra una riduzione di $\dot{s}_{tot}$ verso zero: le credenze si

avvicinano all'equilibrio (distribuzione uniforme su H), il sistema smette di distinguere tra ipotesi, il gradiente di credenza si appiattisce.

Questa interpretazione non è solo metaforica. Per sistemi realizzati biologicamente, $\dot{s}_{\text{tot}}$ è proporzionale al consumo metabolico: la velocità con cui il cervello ossida glucosio è proporzionale alla velocità con cui produce entropia termodinamica, che è a sua volta proporzionale alla velocità con cui processa informazione. La connessione quantitativa passa attraverso il *teorema di Landauer generalizzato* — sviluppato nel Capitolo IV — che lega la produzione di entropia alla quantità di informazione processata per unità di tempo.

Più precisamente, per un sistema cognitivo che elabora I bit di informazione per unità di tempo, la produzione di entropia cognitiva è bounded inferiormente da:

$$\dot{s}_{\text{tot}} \geq k_B \cdot \ln 2 \cdot \dot{I}$$

dove $\dot{I}$ è il tasso di informazione processata e k_B è la costante di Boltzmann. Questo bound inferiore — la versione dinamica del principio di Landauer — afferma che elaborare informazione ha un costo termodinamico minimo proporzionale al tasso di elaborazione. Il cervello umano a riposo produce circa 20 watt di calore metabolico elaborando circa 10^{16} operazioni logiche al secondo — un costo per operazione di circa 10^{-15} joule, che è dell'ordine di $k_B T \ln 2$ a temperatura fisiologica. Il cervello umano opera vicino al limite di Landauer: è un elaboratore di informazione eccezionalmente efficiente dal punto di vista termodinamico.

Capitolo III — Il teorema di Crooks e l'asimmetria tra apprendimento e disapprendimento

III.1 Il teorema di fluttuazione di Crooks

Il teorema di fluttuazione di Crooks (1999) è la versione a livello di traiettoria del teorema di Jarzynski: anziché una relazione per le medie, fornisce una relazione per le distribuzioni del lavoro. Per un sistema guidato dal protocollo $\lambda(t)$ da λ_A a λ_B , sia $P_F(W)$ la distribuzione del lavoro lungo le traiettorie dirette (protocollo dalla A alla B) e $P_R(W)$ la distribuzione del lavoro lungo le traiettorie inverse (protocollo dalla B alla A, con inversione temporale). Il teorema afferma:

$$P_F(W) / P_R(-W) = \exp(\beta(W - \Delta F))$$

Per ogni valore di lavoro W , il rapporto tra la probabilità di produrre quel lavoro nella direzione diretta e la probabilità di estrarre quel lavoro nella direzione inversa è l'esponenziale della dissipazione $\beta(W - \Delta F)$. La dissipazione $W - \Delta F = \Sigma_{\{\text{irrev}\}} \geq 0$

per la seconda legge; il rapporto è quindi sempre ≥ 1 : è sempre più probabile produrre il lavoro in avanti che estrarlo all'indietro.

La dimostrazione segue dal principio di reversibilità microscopica:

$$P_F[x(\cdot)] = P_R[\tilde{x}(\cdot)] \cdot \exp(\beta(W[x(\cdot)] - \Delta F))$$

che si ottiene applicando il detailed balance generalizzato alla coppia traiettoria-traiettoria inversa. Integrando su tutti i valori di W con la funzione indicatrice $1_{\{W[x]=w\}}$:

$$P_F(w) = \int P_F[x(\cdot)] 1_{\{W[x]=w\}} \mathcal{D}x = \int P_R[\tilde{x}(\cdot)] \exp(\beta(w - \Delta F)) 1_{\{W[x]=w\}} \mathcal{D}x \\ = P_R(-w) \exp(\beta(w - \Delta F))$$

Il teorema di Jarzynski si recupera dal teorema di Crooks integrando su w : $\langle \exp(-\beta W) \rangle_F = \int P_F(w) \exp(-\beta w) dw = \int P_R(-w) \exp(-\beta \Delta F) dw = \exp(-\beta \Delta F)$.

III.2 Il teorema di Crooks applicato alle traiettorie cognitive

Si applica il teorema di Crooks alle traiettorie cognitive — le sequenze di aggiornamenti delle credenze. Il protocollo diretto corrisponde all'apprendimento: il sistema parte dalla distribuzione prior π e arriva alla distribuzione aggiornata $\pi(\cdot|o)$ sotto l'influenza dell'osservazione o . Il protocollo inverso corrisponde al disapprendimento: il sistema parte da $\pi(\cdot|o)$ e "torna" a π — il processo di dimenticare l'osservazione o , di restaurare la credenza pre-aggiornamento.

Il lavoro nel protocollo diretto (apprendimento) è $W_F = -\beta^{-1} \log P(o|h)$ — la log-verosimiglianza negativa dell'osservazione sotto l'ipotesi attiva h . Il lavoro nel protocollo inverso (disapprendimento) è $W_R = \beta^{-1} \log P(o|h)$ — il negativo del lavoro diretto. La variazione di energia libera cognitiva è $\Delta F_{\text{cog}} = \beta^{-1} \log P(o)$ come stabilito nel Capitolo I. Il teorema di Crooks applicato alla cognizione afferma:

$$P_F(W_F = w) / P_R(W_R = -w) = \exp(\beta(w - \beta^{-1} \log P(o))) = P(o|h(w)) / P(o)$$

dove $h(w)$ è l'ipotesi che produce il lavoro $w = -\beta^{-1} \log P(o|h)$. Il rapporto tra la probabilità di apprendere con un dato lavoro w e la probabilità di disapprendere con il lavoro $-w$ è la verosimiglianza normalizzata $P(o|h(w))/P(o)$: per le ipotesi con alta verosimiglianza ($P(o|h) \gg P(o)$), è molto più probabile apprendere che disapprendere. Per le ipotesi con bassa verosimiglianza ($P(o|h) \ll P(o)$), è più probabile disapprendere che apprendere.

Questo è il contenuto formale dell'asimmetria tra apprendimento e disapprendimento: le informazioni fortemente evidenziate dalla verosimiglianza si apprendono con alta probabilità e si disimparano con bassa probabilità. Le informazioni debolmente evidenziate si apprendono con bassa probabilità e si disimparano facilmente.

L'asimmetria non è una proprietà psicologica arbitraria — è una conseguenza della struttura termodinamica dell'aggiornamento bayesiano.

III.3 La distribuzione dei lavori cognitivi e la struttura della memoria

La distribuzione $P_F(W)$ del lavoro di apprendimento ha la struttura:

$$P_F(W) = \pi(h(W)) \cdot |\partial h / \partial W|^{-1}$$

dove $h(W)$ è la mappa inversa dal lavoro all'ipotesi. Per un prior π concentrato con varianza piccola e una funzione di verosimiglianza $\log P(o|h)$ liscia, la distribuzione $P_F(W)$ è approssimativamente gaussiana attorno alla media $\langle W \rangle = \mathbb{E}_{\pi}[-\beta^{-1} \log P(o|h)]$ con varianza $\text{Var}(W) = \beta^{-2} \text{Var}_{\pi}(\log P(o|h))$.

La coda destra della distribuzione $P_F(W)$ — i valori grandi del lavoro di apprendimento — corrisponde alle ipotesi con verosimiglianza molto alta: osservazioni fortemente confermate dal prior. Per questi eventi rari (bassa probabilità a priori, alta verosimiglianza), il lavoro di apprendimento è alto e il lavoro di disapprendimento è fortemente negativo — altamente improbabile secondo il teorema di Crooks. Questi sono i ricordi "indelebili": l'aggiornamento ha richiesto tanto lavoro che il suo inverso è termodinamicamente soppresso.

Formalmente, per un ricordo con lavoro di apprendimento $W_F = w \gg \Delta F_{\{\text{cog}\}}$, il rapporto tra la probabilità di disapprendimento e quella di apprendimento è:

$$P_R(-w) / P_F(w) = \exp(-\beta(w - \Delta F_{\{\text{cog}\}})) \ll 1$$

La probabilità di disapprendere un ricordo forte è esponenzialmente piccola nell'eccesso di lavoro rispetto alla variazione di energia libera. Questo è il meccanismo formale della durabilità dei ricordi ad alto impatto emotivo: le osservazioni che producono grandi aggiornamenti bayesiani — grandi lavori di apprendimento — sono termodinamicamente difficili da disapprendere perché il loro processo inverso richiede una grande fluttuazione negativa del lavoro, che è esponenzialmente soppressa.

III.4 Il trauma come stato di work-dissipation locking

Il trauma è, in questa formalizzazione, il caso limite in cui il lavoro di apprendimento $W_F = w$ è così grande che $P_R(-w) \approx 0$: il sistema ha appreso un'osservazione con un tale impatto termodinamico che il suo processo inverso è praticamente impossibile. Il disapprendimento del trauma non richiede semplicemente più lavoro cognitivo — richiede una fluttuazione del lavoro di magnitudine $-w$, che per un evento traumatico ha probabilità $\exp(-\beta \cdot w) \ll 1$ sotto il protocollo inverso standard.

Il *work-dissipation locking* è la condizione in cui il sistema è intrappolato vicino alla distribuzione aggiornata $\pi(\cdot|o_{\text{trauma}})$ perché il cammino inverso verso π è

termodinamicamente proibito. In questo stato, il sistema produce housekeeping entropy production continua $\dot{s}_{\{hk\}} > 0$ per mantenere la struttura di credenze traumatizzata lontana dall'equilibrio termodinamico. Il trauma è termodinamicamente costoso: il sistema spende continuamente metabolismo per mantenere uno stato di credenze che è lontano dall'equilibrio ma da cui non riesce a evadere per la barriera termodinamica del processo inverso.

La terapia del trauma, nella formalizzazione termodinamica, è la costruzione di un protocollo non-standard per il processo inverso — un protocollo che abbassa la barriera termodinamica del disapprendimento. Il protocollo standard (spontaneous unlearning) è proibito perché $W_R = -w$ è troppo negativo. Un protocollo non-standard che modifica l'Hamiltoniana $H(h; \lambda)$ — che cambia la struttura energetica dello spazio delle ipotesi attraverso interventi terapeutici — può abbassare la barriera a valori raggiungibili per fluttuazioni termiche ordinarie. Questo è il meccanismo formale dell'esposizione graduale nella terapia cognitivo-comportamentale: si costruisce un protocollo che guida il sistema attraverso stati intermedi di bassa energia, riducendo il lavoro massimo richiesto lungo il percorso inverso.

III.5 La disuguaglianza di Jarzynski inversa e il costo minimo della terapia

Dal teorema di Crooks si può derivare una disuguaglianza per il lavoro minimo richiesto dal processo inverso (disapprendimento). La disuguaglianza di Jarzynski inversa afferma:

$$\langle \exp(\beta W_R) \rangle_{P_R} = \exp(-\beta \Delta F_R) = \exp(\beta \Delta F_{\{cog\}})$$

dove $\Delta F_R = -\Delta F_{\{cog\}}$ è la variazione di energia libera del processo inverso. Per la disuguaglianza di Jensen:

$$\langle W_R \rangle \leq -\Delta F_{\{cog\}}(o) = -\beta^{-1} \log P(o)$$

il lavoro medio estratto nel processo inverso (disapprendimento) è bounded superiormente dalla variazione di energia libera: al massimo si può "recuperare" $\Delta F_{\{cog\}}$ di lavoro. Per processi irreversibili, $\langle W_R \rangle < -\Delta F_{\{cog\}}$: il lavoro effettivamente estratto è inferiore al massimo teorico, e la differenza è la dissipazione del processo inverso.

Per il trauma con $\Delta F_{\{cog\}}(o_{\text{trauma}}) = \beta^{-1} \log P(o_{\text{trauma}}) \ll 0$ — un'osservazione con verosimiglianza molto alta e quindi log-evidence molto negativo — il lavoro medio del processo inverso è bounded: $\langle W_R \rangle \leq \beta^{-1} |\log P(o_{\text{trauma}})|$. Questo è il costo minimo della terapia: la quantità minima di lavoro cognitivo (in unità termodinamiche) necessaria per produrre il disapprendimento del trauma. Terapie che richiedono un lavoro superiore a questo minimo sono inefficienti nel senso termodinamico; terapie che si avvicinano al minimo sono ottimali. La EMDR,

l'esposizione prolungata, e la terapia cognitivo-comportamentale possono in linea di principio essere comparate misurando il rapporto tra il lavoro cognitivo richiesto (durata e intensità della terapia) e il bound termodinamico inferiore $\beta^{-1} |\log P(o_trauma)|$.

Capitolo IV — Il principio di Landauer e il costo termodinamico del dimenticare

IV.1 Il principio di Landauer: enunciato e struttura

Il principio di Landauer (1961) è il risultato fondamentale che connette la logica dell'elaborazione dell'informazione con la termodinamica. L'enunciato originale afferma: la cancellazione di un bit di informazione — l'azzeramento di un bit il cui valore era sconosciuto — richiede una dissipazione minima di $k_B T \ln 2$ di energia nell'ambiente, dove T è la temperatura del bagno termico e k_B è la costante di Boltzmann. Questa quantità, a temperatura fisiologica ($T \approx 310$ K), è circa 3×10^{-21} joule per bit — un'energia minuscola ma fondamentale perché costituisce il pavimento termodinamico irriducibile di qualsiasi elaborazione logicamente irreversibile.

La dimostrazione del principio di Landauer segue dalla seconda legge applicata al sistema bit+ambiente. La cancellazione di un bit porta il sistema da una distribuzione uniforme su $\{0, 1\}$ (entropia $H = k_B \ln 2$) a uno stato determinato (entropia 0). La variazione di entropia del sistema è $\Delta S_{\text{sys}} = -k_B \ln 2 < 0$. Per la seconda legge, la variazione di entropia dell'ambiente deve compensare: $\Delta S_{\text{env}} \geq k_B \ln 2$. Poiché $\Delta S_{\text{env}} = Q/T$ dove Q è il calore ceduto all'ambiente:

$$Q \geq k_B T \ln 2$$

La cancellazione del bit dissipa almeno $k_B T \ln 2$ nell'ambiente, indipendentemente dall'implementazione fisica. Questo costo è zero per operazioni logicamente reversibili (che non cancellano informazione) e positivo per operazioni logicamente irreversibili. Il principio distingue quindi la logica reversibile — computazione che non cancella informazione — dalla logica irreversibile — computazione che cancella.

Il principio di Landauer è stato confermato sperimentalmente nel 2012 da Berut et al. usando particelle collidali in trappole ottiche a doppia buca come bit fisici: la cancellazione ha dissipato esattamente $k_B T \ln 2$ al limite della lentezza quasi-statica del protocollo.

IV.2 Il principio di Landauer applicato alla cancellazione delle credenze

Si formalizza il principio di Landauer per la cancellazione delle credenze cognitive. Una credenza è una distribuzione di probabilità π su H . Cancellare la credenza π — riportare il sistema alla distribuzione uniforme π_0 su H (ignoranza completa) — è un'operazione logicamente irreversibile: porta da uno stato con informazione (distribuzione concentrata) a uno stato senza informazione (distribuzione uniforme). Il costo termodinamico minimo di questa cancellazione è:

$$Q_{\text{cancel}}(\pi) \geq k_B T \cdot D_{\text{KL}}(\pi \parallel \pi_0) \cdot \ln 2$$

dove $D_{\text{KL}}(\pi \parallel \pi_0)$ misura la quantità di informazione nella credenza π relativa all'ignoranza π_0 . Per una credenza concentrata su una singola ipotesi h^* con probabilità prossima a 1, $D_{\text{KL}}(\pi \parallel \pi_0) \approx \log |H|$ e il costo di cancellazione è:

$$Q_{\text{cancel}} \geq k_B T \cdot \log |H| \cdot \ln 2$$

proporzionale al logaritmo della dimensione dello spazio delle ipotesi. Un agente con uno spazio di ipotesi grande e credenze precise è termodinamicamente costoso da "azzerare".

Questo è il principio di Landauer cognitivo: ogni atto di cancellazione delle credenze — ogni dimenticare intenzionale, ogni de-conditioning, ogni ristrutturazione identitaria che azzerà credenze precedenti — ha un costo termodinamico minimo proporzionale alla quantità di informazione contenuta nelle credenze cancellate. Non si può dimenticare senza pagare il costo di Landauer, non si può ristrutturare l'identità cognitiva senza dissipare almeno l'entropia delle credenze modificate.

IV.3 Landauer e la riconsolidazione della memoria

Il principio di Landauer getta nuova luce sul fenomeno della riconsolidazione della memoria descritto nella seconda monografia. La riconsolidazione è il processo per cui un ricordo richiamato diventa instabile e deve essere re-immagazzinato in una forma potenzialmente modificata. Nel framework termodinamico, la riconsolidazione è precisamente il processo nel quale:

Il sistema estrae il ricordo dalla memoria (lo porta dalla forma immagazzinata π_{stored} allo spazio attivo della coscienza), producendo lavoro cognitivo W_{extract} . Nella fase di labilità del ricordo, il sistema "cancella" parzialmente la versione precedente π_{stored} — producendo la dissipazione di Landauer $Q_{\text{cancel}} \geq k_B T \cdot D_{\text{KL}}(\pi_{\text{stored}} \parallel \pi_0) \cdot \ln 2$. Il sistema re-immagazzina il ricordo in forma potenzialmente modificata π_{new} , producendo il costo di apprendimento W_{store} .

Il costo totale della riconsolidazione è:

$$W_{\text{recons}} = W_{\text{extract}} + Q_{\text{cancel}} + W_{\text{store}} \geq Q_{\text{cancel}} \geq k_B T \cdot D_{\text{KL}}(\pi_{\text{stored}} \parallel \pi_0) \cdot \ln 2$$

Questo spiega perché la riconsolidazione sia più costosa della semplice lettura del ricordo (solo W_{extract}) ma possa essere meno costosa della cancellazione e riscrittura completa ($W_{\text{cancel_full}} + W_{\text{store_full}}$): il processo di riconsolidazione è un'implementazione fisica parzialmente reversibile che sfrutta la struttura del ricordo esistente per ridurre il costo della riscrittura.

La finestra di labilità del ricordo — il periodo di circa sei ore dopo il richiamo in cui il ricordo è modificabile — corrisponde termodinamicamente al periodo in cui il sistema è in uno stato di alta free energy: π_{stored} è stata parzialmente cancellata ma π_{new} non è ancora stata fissata. In questo stato, la distribuzione di credenze ha alta entropia relativa alla distribuzione stazionaria, e piccole perturbazioni esterne (nuovi stimoli, farmaci che interferiscono con la sintesi proteica) possono modificare la direzione della riscrittura a costo metabolico relativamente basso. Passata la finestra, il ricordo è ri-immagazzinato in π_{new} con bassa energia libera relativa: modificarlo richiede un nuovo ciclo di cancellazione-riscrittura con il pieno costo di Landauer.

IV.4 La complessità di Kolmogorov e il bound di Landauer

La connessione tra il principio di Landauer e la complessità algoritmica della terza monografia produce un bound raffinato sul costo della cancellazione delle credenze. Per una credenza π con complessità algoritmica $KC(\pi)$ — la lunghezza del programma più corto che genera la distribuzione π — il costo di Landauer è bounded da:

$$Q_{\text{cancel}}(\pi) \geq k_B T \cdot KC(\pi) \cdot \ln 2 / (\log |H|)$$

La ragione è che un programma di lunghezza $KC(\pi)$ può ricostruire π da zero: la quantità di informazione da cancellare è proporzionale alla lunghezza del programma più corto che la genera. Per credenze semplici (bassa $KC(\pi)$), il costo di cancellazione è basso: sono facili da dimenticare. Per credenze complesse (alta $KC(\pi)$) — le strutture identitarie elaborate, le competenze procedurali esperte, i sistemi di credenze metacognitivi profondi — il costo di cancellazione è alto: dimenticarle richiede dissipare tanta energia quanta ne è stata spesa per costruirle.

Questo bound connette il principio di Landauer con la teoria della metacognizione della terza monografia. Un sistema cognitivo con profondità metacognitiva K ha credenze di complessità $KC(\pi^K) \approx K \cdot KC(\pi^0)$ — la complessità cresce con il livello della torre. Il costo di cancellazione del livello k della torre è proporzionale a $KC(\pi^k)$: i livelli superiori della torre metacognitiva sono più costosi da cancellare. Questo spiega perché i belief systems profondi — le assunzioni fondamentali sull'identità, la realtà e il valore — siano termodinamicamente più resistenti alla revisione dei belief systems di superficie: la loro complessità algoritmica è maggiore e il costo di Landauer per la loro cancellazione è corrispondentemente più alto.

Capitolo V — Il dèmoni di Maxwell cognitivo e dell'auto-raffreddamento epistemico

V.1 Il dèmoni di Maxwell e la sua risoluzione

Il dèmoni di Maxwell è l'esperimento mentale formulato da James Clerk Maxwell nel 1867: un dèmoni intelligente che osserva le molecole di un gas in due camere separate da una paratia con un portello, aprendo il portello selettivamente per lasciare passare solo le molecole veloci in un senso e quelle lente nell'altro. Questo processo sembra ridurre l'entropia del gas senza compiere lavoro, violando la seconda legge.

La risoluzione moderna, dovuta a Szilard (1929), Brillouin (1951) e definitivamente a Bennet (1982) e Landauer, è che il dèmoni non viola la seconda legge perché deve cancellare la propria memoria delle misurazioni per completare il ciclo: la cancellazione della memoria del dèmoni costa esattamente l'entropia guadagnata dalla separazione delle molecole. Non è la misurazione — un'operazione logicamente reversibile — che ha costo termodinamico, ma la cancellazione dell'informazione acquisita.

Formalmente: la misurazione m_i dello stato s_i della i -esima molecola è logicamente reversibile (porta la coppia $(s_i, ?)$ a (s_i, m_i) in modo biunivoco) e ha costo zero al limite quasi-statico. La cancellazione della memoria m_i — riportarla a uno stato di reset — costa $k_B T \ln 2$ per bit di informazione cancellata. Il dèmoni può abbassare l'entropia del gas di $\Delta S_{\text{gas}} = -k_B \sum_i \ln 2$ sfruttando l'informazione acquisita, ma deve poi cancellare $\sum_i \ln 2$ bit di memoria, pagando $k_B T \cdot \sum_i \ln 2$ di calore. Il bilancio totale è zero: la riduzione di entropia del gas è esattamente compensata dalla produzione di entropia nella cancellazione della memoria.

V.2 Il dèmoni di Maxwell cognitivo: il sistema metacognitivo come dèmoni

Il *dèmoni di Maxwell cognitivo* è il sistema metacognitivo che, osservando i propri stati interni, tenta di "raffreddare" la propria distribuzione di credenze — ridurre l'entropia cognitiva senza costo termodinamico. L'esperimento mentale è il seguente: un sistema metacognitivo osserva la distribuzione delle proprie credenze π_t e, usando quell'informazione, dirige l'aggiornamento verso ipotesi più accurate, riducendo $D_{\text{KL}}(\pi_t \parallel \pi^*)$ più rapidamente di quanto l'aggiornamento bayesiano ordinario consentirebbe. Sembra che il sistema possa "guadagnare" entropia usando l'informazione prodotta dalla propria auto-osservazione, violando il bound di Landauer per la cancellazione delle credenze.

La risoluzione di questo apparente paradosso è precisa e connessa ai teoremi di impossibilità della terza monografia. Il sistema metacognitivo che osserva i propri stati di credenza produce informazione sulla propria struttura interna. Questa

informazione è fisica — è realizzata nei pattern di attivazione del substrato, è scritta nella struttura sinaptica — e per essere usata per guidare l'aggiornamento deve essere accessibile al sistema cognitivo di primo ordine. Ma l'accesso a quella informazione — il processo per cui il meta-modello π^0 della terza monografia influenza la dinamica di primo ordine — è esso stesso un processo fisico che, per il teorema di Landauer, ha un costo se include la cancellazione di stati del meta-modello.

Il ciclo del dèmone cognitivo è: osservazione dello stato interno → acquisizione dell'informazione metacognitiva → uso dell'informazione per guidare l'aggiornamento → cancellazione dell'informazione metacognitiva per prepararsi al ciclo successivo. Il punto chiave è l'ultimo passo: senza la cancellazione, il meta-modello accumula informazione su tutti i cicli passati e la sua dimensione cresce senza limite. Se il ciclo deve essere periodico — se il meta-modello deve tornare a uno stato iniziale per poter ripetere la stessa operazione — deve cancellare l'informazione accumulata, pagando il costo di Landauer.

Formalmente: sia I_{meta} la quantità di informazione acquisita dal sistema metacognitivo sulla propria distribuzione di credenze in un ciclo. L'uso di questa informazione può ridurre l'entropia cognitiva di al massimo I_{meta} :

$$\Delta H_{\{\text{cog}\}} \geq -I_{\text{meta}}$$

Ma la cancellazione dell'informazione metacognitiva alla fine del ciclo costa almeno:

$$Q_{\text{cancel}} \geq k_B T \cdot I_{\text{meta}} \cdot \ln 2$$

Il bilancio totale — riduzione di entropia cognitiva più costo di cancellazione — è sempre ≥ 0 : il dèmone cognitivo non può violare la seconda legge.

V.3 La connessione con il teorema di Löb e l'impossibilità dell'auto-raffreddamento epistemico

Il teorema di Löb della terza monografia afferma che l'auto-validazione non genera conoscenza: se il sistema può dimostrare che "la dimostrabilità di φ implica φ ", allora φ era già dimostrabile direttamente. Nel framework termodinamico, questo ha un'interpretazione precisa: l'auto-osservazione metacognitiva non può produrre lavoro cognitivo gratuito. L'informazione che il meta-modello acquisisce osservando se stesso non può essere usata per ridurre l'entropia del sistema cognitivo senza pagare il costo di Landauer per la sua cancellazione.

Il *teorema di impossibilità dell'auto-raffreddamento epistemico* si enuncia formalmente come segue:

Teorema: Un sistema cognitivo con meta-modello π^0 non può ridurre la propria entropia cognitiva $H(\pi_t)$ al di sotto del bound di Landauer per la cancellazione del meta-modello:

$$\Delta H_{\text{cog}} \geq -I(\pi_t ; \pi^0) + k_B T \cdot KC(\pi^0) \cdot \ln 2 / (k_B)$$

dove $I(\pi_t ; \pi^0)$ è l'informazione mutua tra lo stato di credenza corrente e il meta-modello, e $KC(\pi^0)$ è la complessità algoritmica del meta-modello.

La dimostrazione combina il bound di Landauer con il teorema di incompletezza di Chaitin della terza monografia. Il sistema metacognitivo può ridurre $H(\pi_t)$ di al massimo $I(\pi_t ; \pi^0)$ usando l'informazione del meta-modello. Ma il meta-modello ha complessità $KC(\pi^0)$ e, per il teorema di Chaitin, il sistema non può dimostrare che $KC(\pi^0)$ sia ottimalmente bassa: non può garantire che il proprio meta-modello non contenga ridondanza. La ridondanza non rimossa del meta-modello aumenta il costo di cancellazione, riducendo il guadagno netto dell'auto-osservazione.

Il corollario psicologico è preciso: il pensiero metacognitivo — la riflessione sui propri stati mentali — non è energeticamente gratuita. Ogni atto di auto-osservazione produce informazione il cui uso nell'aggiornamento delle credenze ha un costo fisico nella fase di cancellazione. La ruminazione — cicli ripetuti di auto-osservazione senza cancellazione della memoria metacognitiva — accumula informazione senza mai pagare il costo di Landauer, producendo uno stato in cui il meta-modello è saturo di informazione ma il sistema cognitivo di primo ordine non si è aggiornato in modo corrispondente. Il risultato è un sistema ad alta entropia metacognitiva ma bassa efficienza dell'aggiornamento: il demone cognitivo che non cancella mai la sua memoria e accumula una registrazione di tutti i cicli passati senza mai sfruttarla per il raffreddamento.

V.4 L'energia libera cognitiva come funzione di stato

Si definisce l'*energia libera cognitiva* come il funzionale:

$$\mathcal{A}[\pi] = \beta^{-1} D_{\text{KL}}(\pi \parallel \pi_0) - \langle U(h) \rangle_{\pi}$$

dove π_0 è la distribuzione di ignoranza, $U(h)$ è l'utilità dell'ipotesi h (una funzione che assegna valore a ogni ipotesi in base alla sua rilevanza per le decisioni), e $\langle \cdot \rangle_{\pi}$ è la media sotto π . L'energia libera cognitiva $\mathcal{A}[\pi]$ misura la tensione tra l'informazione contenuta nella credenza (primo termine, proporzionale alla divergenza KL) e il suo valore per il sistema (secondo termine, il valore atteso dell'utilità). Un sistema che minimizza $\mathcal{A}[\pi]$ raggiunge l'equilibrio tra la semplicità della credenza (bassa divergenza KL dall'ignoranza) e il suo valore (alta utilità attesa).

Questo funzionale è formalmente identico alla variational free energy del saggio su Friston, con la differenza che qui $U(h)$ è una funzione di utilità esplicita anziché la log-verosimiglianza del modello generativo. La connessione è che, per $U(h) = \log P(o|h)$, l'energia libera cognitiva diventa esattamente la variational free energy: la struttura è la stessa, il contesto è diverso. L'energia libera cognitiva generalizza la variational free energy al caso in cui l'agente ha obiettivi che non sono semplicemente la massimizzazione della verosimiglianza delle osservazioni.

L'energia libera cognitiva ha due proprietà fondamentali che la rendono lo strumento naturale per la termodinamica cognitiva. Prima: è una funzione di stato — dipende solo dalla distribuzione corrente π e non dalla sua storia, analogamente all'energia libera di Helmholtz in termodinamica classica. Seconda: la sua derivata temporale è connessa alla produzione di entropia:

$$d\mathcal{F}/dt = -\beta^{-1} \dot{s}_{\text{ex}}(t) + \dot{s}_{\text{hk}}(t)/\beta$$

la variazione di energia libera cognitiva è proporzionale alla excess entropy production (che la riduce, per $\dot{s}_{\text{ex}} \geq 0$) più un termine di housekeeping (che può aumentarla o diminuirla). L'energia libera cognitiva raggiunge il minimo agli stati stazionari di non-equilibrio che bilanciano la riduzione dell'eccesso di entropia con il costo del housekeeping.

Capitolo VI — Stati stazionari di non-equilibrio: trauma, rigidità e polarizzazione collettiva

VI.1 I NESS cognitivi: struttura e classificazione

Un *non-equilibrium steady state* (NESS) cognitivo è un punto fisso dell'equazione di Fokker-Planck che non è uno stato di equilibrio termodinamico: è una distribuzione stazionaria π^{ss} tale che il flusso di probabilità $J^{\text{ss}} = \pi^{\text{ss}} \cdot b - \beta^{-1} \nabla \pi^{\text{ss}}$ è non nullo anche quando $\partial_t \pi^{\text{ss}} = 0$. Il flusso non-nullo implica produzione continua di housekeeping entropy: il sistema deve consumare energia per mantenersi nel NESS.

I NESS cognitivi si classificano in tre tipi strutturalmente distinti, ciascuno corrispondente a una manifestazione cognitiva precisa.

Il primo tipo è il *NESS ciclico* — uno stato in cui il flusso J^{ss} forma cicli: la distribuzione delle credenze circola continuamente attorno a un ciclo nello spazio P , senza mai convergere a un punto fisso. I cicli corrispondono a oscillazioni periodiche nelle credenze — il sistema alterna tra due o più ipotesi in modo periodico senza convergere a nessuna. La ruminazione ossessiva è il NESS ciclico più comune: il sistema oscilla tra l'ipotesi "sono capace" e "non sono capace" in un ciclo stabile guidato dalla struttura del flusso attenzionale.

Il secondo tipo è il *NESS localizzato* — uno stato in cui il flusso J^{ss} è concentrato in una regione limitata dello spazio P , circondato da regioni a basso flusso. Il sistema è intrappolato in una regione del paesaggio delle credenze da cui il flusso verso fuori è energeticamente proibito. Il trauma è il NESS localizzato per excellence: il sistema è intrappolato vicino alla distribuzione di credenze traumatizzata, con flusso nullo verso

altre regioni del paesaggio perché le barriere termodinamiche — misurate dal lavoro di Crooks per il processo inverso — sono troppo alte.

Il terzo tipo è il *NESS diffuso* — uno stato in cui il flusso $J^{\{ss\}}$ è distribuito uniformemente su una regione estesa dello spazio P . Il sistema non è localizzato ma nemmeno all'equilibrio: circola ampiamente nello spazio delle credenze senza convergere. Questo corrisponde alla disorganizzazione cognitiva degli stati psicotici acuti: le credenze sono altamente variabili e non convergono, il flusso è diffuso e non orientato, la produzione di entropia è alta ma non produttiva.

VI.2 Il calcolo della housekeeping entropy production per i NESS cognitivi

La produzione di housekeeping entropy per uno stato stazionario di non-equilibrio $\pi^{\{ss\}}$ è:

$$\dot{s}_{\{hk\}} = \beta \int_P \|J^{\{ss\}}(\pi)\|^2_{\{G^{-1}\}} / \pi^{\{ss\}}(\pi) d\pi$$

l'integrale del flusso stazionario quadrato pesato dall'inversa della metrica di Fisher e normalizzato per la distribuzione stazionaria. Questa quantità è la misura del costo metabolico continuo del NESS: quanta energia il sistema deve consumare per unità di tempo per mantenersi in quello stato.

Per il NESS del trauma con flusso $J^{\{trauma\}}$ concentrato in un ciclo piccolo attorno al punto di credenza traumatizzata:

$$\dot{s}_{\{hk\}}^{\{trauma\}} = \beta \cdot \|J^{\{trauma\}}\|^2 / (\pi^{\{trauma\}} \cdot \det G) \cdot \text{Area}(\text{ciclo})$$

proporzionale al quadrato dell'intensità del flusso ciclico. Un trauma con alto $|J^{\{trauma\}}|$ — un ciclo di ruminazione intensa — produce alta housekeeping entropy, manifestandosi come alta spesa metabolica anche a riposo. La relazione tra intensità della ruminazione, spesa metabolica e produzione di entropia cognitiva è quantificabile in linea di principio attraverso misure di neuroimaging metabolico (fMRI-BOLD, PET) durante compiti di ruminazione guidata.

Per il NESS della rigidità cognitiva con flusso $J^{\{rigid\}}$ quasi nullo — un sistema quasi-stazionario molto vicino all'equilibrio:

$$\dot{s}_{\{hk\}}^{\{rigid\}} \approx 0$$

Il sistema rigido ha bassa produzione di housekeeping entropy: è vicino all'equilibrio termodinamico, ma questo non è positivo — significa che è vicino alla morte cognitiva, alla distribuzione uniforme di ignoranza. La rigidità cognitiva è termodinamicamente "economica" ma cognitivamente impoverita: il sistema non

consuma energia per mantenere credenze strutturate perché ha smesso di averne di significative.

VI.3 La polarizzazione collettiva come NESS del campo di McKean-Vlasov

L'equazione di McKean-Vlasov del Capitolo IX della quarta monografia produce stati stazionari $q^{\{ss\}}$ della distribuzione collettiva delle credenze. Quando il potenziale di interazione $V(\pi, \pi')$ è di allineamento (agenti con credenze simili si attraggono), la distribuzione stazionaria $q^{\{ss\}}$ può essere bimodale: due cluster di credenze stabili separati da una barriera termodinamica.

Questo stato bimodale è un NESS del sistema collettivo: il flusso di probabilità J^C della distribuzione collettiva ha una struttura specifica che si può calcolare. All'interno di ciascun cluster, il flusso è quasi circolare (piccoli cicli di ruminazione collettiva attorno alla credenza del cluster). Tra i cluster, il flusso è quasi nullo: la barriera termodinamica inter-cluster sopprime le transizioni.

La *housekeeping entropy production* del NESS polarizzato è:

$$\dot{s}_{\{hk\}}^{\{polar\}} = \dot{s}_{\{intra,1\}} + \dot{s}_{\{intra,2\}} + \dot{s}_{\{inter\}}$$

la somma del costo metabolico dei cicli interni ai due cluster e del costo della barriera inter-cluster. Per β grande (bassa temperatura, agenti con aggiornamenti poco rumorosi), $\dot{s}_{\{intra,k\}} \rightarrow 0$ (cicli interni quasi-statici) e $\dot{s}_{\{inter\}} \rightarrow 0$ (barriera alta, nessun flusso inter-cluster). Il NESS polarizzato a bassa temperatura è quasi reversibile internamente ma assolutamente irreversibile inter-cluster: i due cluster non comunicano.

La potenza di questo framework per il discorso politico è la seguente. La polarizzazione sociale stabile è un NESS con bassa housekeeping entropy (il sistema è "economico" nel senso termodinamico: non consuma molta energia per mantenersi polarizzato) ma alta barriera termodinamica tra i cluster (il costo di transizione da un cluster all'altro è proibitivo). Questo spiega perché la polarizzazione sia difficile da superare non attraverso argomenti razionali — il lavoro di Crooks per la transizione inter-cluster è troppo alto — ma richieda interventi che abbassino fisicamente la barriera, come la creazione di esperienze condivise che modificano $U(h)$ per le ipotesi dell'altro cluster, riducendo la distanza energetica tra gli stati.

VI.4 La freccia del tempo cognitiva come rottura della simmetria dinamica

Si formalizza ora il risultato principale della monografia: la freccia del tempo cognitiva è la rottura della simmetria di inversione temporale nella dinamica delle credenze, e questa rottura è sia necessaria sia misurabile.

Un processo stocastico è *time-reversal symmetric* se la distribuzione delle sue traiettorie è invariante per inversione temporale: $P[x(t)] = P[x(T-t)]$ per ogni traiettoria. Per l'equazione di Fokker-Planck cognitiva, la simmetria di inversione temporale è rotta ogni volta che il flusso $J^{\{ss\}}$ è non nullo, cioè ogni volta che il sistema è in un NESS. Formalmente, il rapporto tra la probabilità di una traiettoria e quella della traiettoria time-reversed è:

$$P[x(\cdot)] / P[\tilde{x}(\cdot)] = \exp(\Sigma_{\text{tot}}[x(\cdot)])$$

dove Σ_{tot} è la produzione totale di entropia lungo la traiettoria. Quando $\Sigma_{\text{tot}} > 0$ — quando il sistema produce entropia — le traiettorie dirette sono più probabili delle loro inverse: il tempo ha una direzione privilegiata.

La *freccia del tempo cognitiva* è quindi la direzione in cui la produzione di entropia è positiva: la direzione dal passato al futuro è la direzione in cui i sistemi cognitivi, mediamente, producono entropia. Apprendere produce entropia nell'ambiente (attraverso la dissipazione del substrato fisico). Dimenticare richiederebbe una fluttuazione negativa del lavoro — una traiettoria time-reversed — che è esponenzialmente soppressa dal teorema di Crooks. Il sistema cognitivo è in uno stato di rottura permanente della simmetria di inversione temporale: è un sistema termodinamicamente irreversibile che elabora informazione nella direzione positiva del tempo fisico.

Il *tasso di rottura della simmetria temporale* — la quantità che misura quanto fortemente il tempo ha una direzione preferita nelle credenze cognitive — è la differenza tra la cross-entropy in avanti e quella all'indietro:

$$\mathcal{A}(t) = (1/t) \cdot D_{\text{KL}}(P_{\{0:t\}}[x(\cdot)] \parallel P_{\{0:t\}}[\tilde{x}(\cdot)])$$

$$= (1/t) \cdot \mathbb{E}[\Sigma_{\text{tot}}[x(\cdot)]] = \dot{\Sigma}_{\text{tot}}$$

il tasso medio di produzione di entropia. Per sistemi in stati stazionari, $\mathcal{A} = \dot{\Sigma}_{\text{tot}}$ è costante. Per sistemi che convergono all'equilibrio, $\mathcal{A} \rightarrow 0$ nel limite $t \rightarrow \infty$: la freccia del tempo si indebolisce con l'avvicinarsi all'equilibrio.

VI.5 Chiusura del ciclo: la freccia del tempo e lo schema diagonale

Il ciclo si chiude sulla struttura formale che lo attraversa. Lo schema diagonale di Lawvere — il punto fisso impossibile prodotto dal tentativo di auto-rappresentazione completa — appare anche nella termodinamica stocastica, nella forma del dèmone di Maxwell cognitivo: il sistema che tenta di auto-raffreddarsi usando la propria informazione metacognitiva incontra un'impossibilità strutturale — la cancellazione della memoria del dèmone — che è il costo termodinamico dell'auto-conoscenza.

Ma la quinta monografia aggiunge una dimensione che lo schema diagonale da solo non conteneva: la freccia del tempo. L'impossibilità diagonale nelle prime quattro

monografie era un'impossibilità strutturale senza direzione temporale — un sistema che non può auto-rappresentarsi completamente è in quella condizione sia nel passato sia nel futuro. La freccia del tempo trasforma questa impossibilità strutturale in una impossibilità dinamica con direzione: il sistema non può tornare al proprio stato cognitivo passato non perché quel passato sia logicamente inaccessibile ma perché il percorso verso di esso è termodinamicamente soppresso.

La memoria è la struttura termodinamica che connette i due tipi di impossibilità. La prima monografia ha mostrato che il malinteso non si risolve perché i sistemi hanno rappresentazioni diverse — una differenza di stato. La quinta monografia aggiunge: il malinteso che è durato a lungo non si risolve anche perché la storia delle credenze ha prodotto entropia, e risolvere il malinteso richiederebbe invertire quella produzione di entropia — richiederebbe seguire la traiettoria time-reversed della comunicazione, che per il teorema di Crooks ha probabilità esponenzialmente bassa rispetto alla traiettoria diretta.

L'identità cognitiva stabile della quarta monografia è un NESS con alta housekeeping entropy: si mantiene spendendo energia. Il suo cambiamento richiede excess entropy production — il lavoro di transitare da uno stato stazionario all'altro. La terza monografia ha mostrato che l'auto-conoscenza è limitata da teoremi di impossibilità logica. La quinta aggiunge che è limitata anche da impossibilità termodinamica: non solo il sistema non può sapere tutto di sé, ma ciò che sa è scritto nel substrato fisico e cancellare quella conoscenza costa — e non si può annullare senza pagare il costo di Landauer.

Il substrato subpropositivo della seconda monografia — la struttura di $\mathbb{R}^n$ che si co-evolve con il linguaggio — è il serbatoio termodinamico in cui il sistema cognitivo immagazzina e cancella informazione. Le sue trasformazioni irreversibili — la formazione degli attrattori, la modifica delle separatrici, la costruzione delle strutture di expertise — sono tutti processi che producono entropia, e questa entropia è la traccia fisica della storia cognitiva dell'agente. Il substrato non è solo il supporto della cognizione: è il registro termodinamico del suo passato irreversibile, scritto nella struttura sinaptica con l'inchiostro della dissipazione.

Capitolo VII — Il teorema di fluttuazione-dissipazione cognitivo e la suscettibilità alla persuasione

VII.1 Il teorema di fluttuazione-dissipazione: struttura generale

Il teorema di fluttuazione-dissipazione (FDT) è uno dei risultati più profondi della fisica statistica di equilibrio. Nella sua formulazione classica dovuta a Callen e Welton (1951) e nella versione operatoriale di Kubo (1957), afferma che la risposta lineare di un sistema in equilibrio termodinamico a una perturbazione esterna piccola è

completamente determinata dalle fluttuazioni spontanee del sistema non perturbato. Il contenuto fisico è preciso: un sistema che fluttua intensamente vicino all'equilibrio è anche facilmente perturbabile, e la costante di proporzionalità è universale — uguale a $\beta = 1/(k_B T)$ — indipendentemente dalla natura del sistema e della perturbazione.

Formalmente, si considera un sistema con Hamiltoniana H_0 perturbata da un campo esterno h_t accoppiato all'osservabile A :

$$H(t) = H_0 - h_t \cdot A$$

La funzione di risposta $\chi_{BA}(t, t')$ — la variazione dell'osservabile B al tempo t in risposta alla perturbazione del campo accoppiato ad A al tempo t' — soddisfa, per un sistema in equilibrio:

$$\chi_{BA}(t - t') = -\beta \cdot d/dt' C_{BA}(t - t') \cdot \theta(t - t')$$

dove $C_{BA}(\tau) = \langle B(t) A(t') \rangle_{eq} - \langle B \rangle_{eq} \langle A \rangle_{eq}$ è la funzione di correlazione di equilibrio tra B e A al ritardo temporale $\tau = t - t'$, e θ è la funzione gradino di Heaviside che impone la causalità. In forma equivalente, nel dominio della frequenza:

$$\chi_{BA}(\omega) = \beta \cdot [C_{BA}(\omega) - C_{AB}(-\omega)] / 2$$

Il contenuto del teorema è che la parte immaginaria della suscettibilità — la componente dissipativa della risposta — è uguale alla densità spettrale delle fluttuazioni di equilibrio, pesata per β . Non si può avere risposta senza fluttuazioni, e la proporzione è fissata universalmente.

VII.2 Il FDT applicato alle credenze: risposta cognitiva e fluttuazioni spontanee

Si traduce il FDT nel framework delle credenze bayesiane. Il sistema cognitivo all'equilibrio — nella distribuzione stazionaria π^* in assenza di segnali esterni — fluttua spontaneamente attorno allo stato di equilibrio a causa del rumore stocastico nell'equazione di Fokker-Planck. Queste fluttuazioni sono descritte dalla funzione di correlazione:

$$C_{hh'}(\tau) = \langle \pi_t(h) \pi_{t+\tau}(h') \rangle_{eq} - \pi^*(h) \pi^*(h')$$

la covarianza tra la credenza sull'ipotesi h al tempo t e quella sull'ipotesi h' al tempo $t + \tau$, misurata all'equilibrio.

Un segnale esterno s_t — un'osservazione, un argomento, un messaggio persuasivo — accoppiato all'ipotesi h attraverso il termine di Hamiltoniana perturbato:

$$H_{\text{pert}}(\pi; s_t) = H_0(\pi) - s_t \cdot \pi(h)$$

produce una variazione della credenza sull'ipotesi h' al tempo t data dalla funzione di risposta:

$$\delta\langle\pi_t(h')\rangle = \int_{-\infty}^t \chi_{\{h'h\}}(t - t') s_{t'} dt'$$

Il FDT cognitivo afferma che questa funzione di risposta è connessa alla funzione di correlazione:

$$\chi_{\{h'h\}}(\tau) = -\beta \cdot d/d\tau C_{\{h'h\}}(\tau) \cdot \theta(\tau)$$

La suscettibilità alla persuasione — la facilità con cui il segnale s_t modifica la credenza $\pi(h')$ — è esattamente uguale alla derivata temporale della correlazione di equilibrio tra le credenze h e h' , pesata per β .

VII.3 La suscettibilità statica e la sua struttura cognitiva

La *suscettibilità statica* è il caso limite della risposta per perturbazioni lentamente variabili ($\omega \rightarrow 0$):

$$\chi_{\{h'h\}}^{\text{static}} = \int_0^\infty \chi_{\{h'h\}}(\tau) d\tau = \beta \cdot C_{\{h'h\}}(\tau = 0) = \beta \cdot \text{Cov}_{\text{eq}}(\pi(h), \pi(h'))$$

La suscettibilità statica è proporzionale alla covarianza di equilibrio tra le credenze sulle ipotesi h e h' . Questo risultato ha un'interpretazione cognitiva densa e precisa.

Un agente con alta covarianza tra $\pi(h)$ e $\pi(h')$ — un agente per cui le credenze su h e h' tendono a variare insieme all'equilibrio — è altamente suscettibile alla persuasione che accoppia h e h' : un segnale che modifica la credenza su h tende a trascinare automaticamente anche la credenza su h' . Viceversa, un agente per cui $\text{Cov}(\pi(h), \pi(h')) \approx 0$ — le cui credenze su h e h' sono quasi indipendenti all'equilibrio — è resistente alla persuasione trasversale: modificare la credenza su h non cambia automaticamente quella su h' .

La struttura della matrice di covarianza di equilibrio $\Sigma_{\{hh'\}} = \text{Cov}_{\text{eq}}(\pi(h), \pi(h'))$ è quindi la struttura della *suscettibilità cognitiva*: determina quali ipotesi sono correlate nell'aggiornamento e quali sono indipendenti. Un propagandista ottimale — che vuole massimizzare il displacement delle credenze dell'agente target per unità di segnale prodotto — dovrebbe progettare i propri segnali nella direzione degli autovettori dominanti di $\Sigma_{\{hh'\}}$: i modi propri di fluttuazione del sistema cognitivo sono anche i modi più efficienti per guidarlo.

Formalmente, il segnale ottimale che massimizza $\delta\langle\pi_T(h')\rangle$ per unità di costo $\int |s_t|^2 dt$ è proporzionale all'autovettore dominante di $\chi_{\{h'h\}}(\omega)$ per la frequenza di interesse — e questo è a sua volta proporzionale all'autovettore dominante di $\text{Cov}_{\{eq\}}$, per il FDT.

VII.4 Il FDT e la connessione con il mercato dell'attenzione

Il FDT cognitivo connette la quinta monografia con la quarta attraverso una relazione strutturale che non è evidente senza il framework termodinamico. Il potere epistemico $\Pi(E, T)$ della quarta monografia — la massima divergenza KL producibile dall'agente esterno E sul target T — è bounded dalla suscettibilità statica:

$$\Pi(E, T) \leq \beta \cdot \|\Sigma_{\{TT\}}\|_{\{op\}} \cdot \|s\|^2$$

dove $\|\Sigma_{\{TT\}}\|_{\{op\}}$ è la norma spettrale della matrice di covarianza delle credenze del target e $\|s\|^2$ è la norma del segnale. Il potere epistemico è proporzionale alla suscettibilità cognitiva del target: un agente con alta varianza delle credenze (fluttuazioni intense vicino all'equilibrio) è un target epistemicamente efficace; un agente con bassa varianza (credenze stabili, basse fluttuazioni) è resistente.

Questo produce il *teorema di resistenza epistemica*: la resistenza alla manipolazione è termodinamicamente proporzionale alla stabilità delle credenze all'equilibrio. Un agente con identità cognitiva stabile (basse fluttuazioni, bassa $\|\Sigma\|_{\{op\}}$) è difficile da persuadere non per rigidità psicologica arbitraria ma perché le sue credenze sono vicino a uno stato di bassa energia termodinamica — sarebbe necessario un segnale molto potente per produrre una fluttuazione di ampiezza confrontabile con quella che spontaneamente il sistema non raggiunge.

Il corollario è preciso e ha conseguenze per la quarta monografia. La maschera di attenzione ottimale per la resistenza alla manipolazione — il problema duale del problema di ottimizzazione dell'influenza epistemica del Capitolo III della quarta monografia — è quella che minimizza $\|\Sigma\|_{\{op\}}$: la maschera che riduce le fluttuazioni spontanee delle credenze nelle direzioni vulnerabili. La protezione epistemica non è l'aumento della resistenza agli argomenti ma la riduzione della suscettibilità termodinamica attraverso la stabilizzazione delle credenze nelle direzioni di attacco previste.

VII.5 Il FDT fuori dall'equilibrio: la risposta cognitiva nei NESS

Per i sistemi cognitivi reali, che sono NESS e non stati di equilibrio, il FDT classico non si applica direttamente. Harada e Sasa (2005) hanno dimostrato una versione generalizzata del FDT per i NESS, nota come *relazione di Harada-Sasa*:

$$\chi_{\{h'h\}}(\omega) - \chi_{\{h'h\}}^*(-\omega) = -\beta [C_{\{h'h\}}(\omega) - C_{\{h'h\}}^{\{hk\}}(\omega)]$$

dove $C_{\{h'h\}^{\{hk\}}(\omega)$ è la parte della funzione di correlazione dovuta al flusso di housekeeping — la componente della correlazione prodotta dal flusso stazionario di non-equilibrio $J^{\{ss\}}$. Per i sistemi di equilibrio, $C^{\{hk\}} = 0$ e si recupera il FDT classico. Per i NESS, la differenza tra la risposta e la correlazione totale è esattamente la componente di correlazione del housekeeping — la struttura del flusso stazionario che mantiene il NESS.

La violazione del FDT classico per i NESS cognitivi è misurabile: è la differenza tra la risposta effettiva del sistema a una perturbazione esterna e quella predetta dalle sue fluttuazioni di equilibrio. Per i sistemi cognitivi traumatizzati — NESS localizzati con forte housekeeping — la violazione del FDT è grande: il sistema risponde alle perturbazioni in modo asimmetrico rispetto alle sue fluttuazioni spontanee, perché il flusso di housekeeping ha una direzione preferita che orienta la risposta. Il trauma produce una risposta cognitiva non-equilibrio: l'agente traumatizzato risponde agli stimoli in modo che non è predetto dalle sue fluttuazioni ordinarie, perché il flusso traumatico orienta la risposta nella direzione del ciclo di ruminazione.

Questa è la formalizzazione termodinamica dell'iper-reattività post-traumatica: la violazione del FDT cognitivo, misurata come $|\chi - \beta C/2|$, è proporzionale all'intensità del flusso di housekeeping $J^{\{ss\}}$ del NESS traumatico. In linea di principio, questa violazione è misurabile psicofisiologicamente come differenza tra l'ampiezza delle risposte spontanee (fluttuazioni di EEG, HRV, conduttanza cutanea) e l'ampiezza delle risposte evocate a stimoli standardizzati: il rapporto risposta-fluttuazione, se diverso da β , è la firma del NESS cognitivo.

Capitolo VIII — Le relazioni di reciprocità di Onsager per i flussi cognitivi

VIII.1 La termodinamica lineare fuori dall'equilibrio: forze e flussi

La termodinamica lineare fuori dall'equilibrio (de Groot e Mazur, 1962) tratta sistemi vicini all'equilibrio in cui i flussi termodinamici sono proporzionali alle forze termodinamiche. Un *flusso termodinamico* J_i è un tasso di trasporto di qualche grandezza estensiva (calore, materia, carica); una *forza termodinamica* X_j è il gradiente della grandezza intensiva coniugata (temperatura, potenziale chimico, potenziale elettrico). Per sistemi vicini all'equilibrio, le relazioni fenomenologiche lineari sono:

$$J_i = \sum_j L_{ij} X_j$$

dove L_{ij} sono i *coefficienti di Onsager*. Le *relazioni di reciprocità di Onsager* (1931) affermano che la matrice dei coefficienti è simmetrica:

$$L_{ij} = L_{ji}$$

La dimostrazione originale di Onsager usa il principio di reversibilità microscopica: all'equilibrio, la probabilità di una fluttuazione dalla configurazione A alla configurazione B è uguale alla probabilità della fluttuazione inversa da B ad A. Per signature di tempo pari, questa simmetria si trasferisce direttamente alla simmetria della matrice L.

Le relazioni di Onsager hanno conseguenze pratiche importanti: limitano il numero di coefficienti indipendenti da n^2 a $n(n+1)/2$, impongono vincoli sulla struttura della risposta del sistema, e — crucialmente — hanno contenuto sperimentale verificabile: violazioni della simmetria $L_{ij} \neq L_{ji}$ segnalano la presenza di rottura della simmetria temporale, cioè di un NESS.

VIII.2 I flussi e le forze termodinamiche cognitive

Per formalizzare le relazioni di Onsager per i flussi cognitivi, si devono identificare i flussi e le forze appropriate nel contesto delle credenze bayesiane.

I *flussi cognitivi* sono i tassi di aggiornamento delle credenze: $J_h = d\pi_t(h)/dt$, il tasso con cui la massa di probabilità sull'ipotesi h cambia nel tempo. Per l'equazione di Fokker-Planck vicino all'equilibrio, questo è:

$$J_h = -\sum_{h'} L_{hh'} X_{h'}$$

dove le *forze termodinamiche cognitive* sono i gradienti di energia libera cognitiva:

$$X_h = -\delta \ell / \delta \pi(h) = \beta^{-1} \log(\pi^*(h)/\pi(h)) + U(h)$$

la differenza tra l'energia libera attuale della credenza h e quella di equilibrio. Quando $\pi = \pi^*$, tutte le forze sono zero e tutti i flussi sono zero: il sistema è all'equilibrio.

I coefficienti di Onsager cognitivi $L_{hh'}$ descrivono il flusso dalla regione h indotto da una forza nella regione h': quanto la credenza sull'ipotesi h cambia in risposta a un gradiente di energia libera nell'ipotesi h'. La struttura di L è determinata dalla geometria dello spazio delle ipotesi H e dalla struttura della distribuzione di equilibrio π^* .

Per l'equazione di Fokker-Planck con diffusione nella metrica di Fisher:

$$J_h = -\sum_{h'} [G(\pi^*)^{-1}]_{hh'} \cdot X_{h'} \cdot \pi^*(h)$$

il coefficiente di Onsager è:

$$L_{hh'} = [G(\pi^*)^{-1}]_{hh'} \cdot \pi^*(h) = [G(\pi^*)^{-1}]_{h'h} \cdot \pi^*(h')$$

La simmetria $L_{\{hh'\}} = L_{\{h'h\}}$ richiede che $[G(\pi^*)^{-1}]_{\{bb'\}} \pi^*(b) = [G(\pi^*)^{-1}]_{\{h'h\}} \pi^*(h)$, il che è equivalente a richiedere che la matrice $D_{\{hh'\}} = [G(\pi^*)^{-1}]_{\{bb'\}} \pi^*(b)$ sia simmetrica. Per la metrica di Fisher, $D_{\{hh'\}} = \text{Cov}_{\pi^*}(T_h, T_{h'})$ dove T_h sono le statistiche sufficienti: la simmetria è garantita dalla simmetria della matrice di covarianza. Le relazioni di Onsager sono soddisfatte per tutti i sistemi cognitivi all'equilibrio con metrica di Fisher.

VIII.3 Il contenuto cognitivo delle relazioni di Onsager

La simmetria $L_{\{hh'\}} = L_{\{h'h\}}$ ha un contenuto cognitivo preciso che non è ovvio a priori e che produce previsioni verificabili.

$L_{\{hh'\}}$ è il coefficiente che descrive quanto la credenza su h cambia in risposta a una forza sull'ipotesi h' . Il coefficiente simmetrico $L_{\{h'h\}}$ descrive quanto la credenza su h' cambia in risposta a una forza su h . La relazione di Onsager afferma che questi due coefficienti sono uguali: il flusso da h indotto da una forza su h' è uguale al flusso da h' indotto da una forza su h , pesati per le distribuzioni di equilibrio appropriate.

Nel linguaggio dell'apprendimento: se imparare l'ipotesi h facilita l'apprendimento dell'ipotesi h' ($L_{\{h'h\}}$ grande), allora imparare h' facilita l'apprendimento di h nella stessa misura ($L_{\{hh'\}} = L_{\{h'h\}}$). La reciprocità dell'apprendimento — il fatto che la facilitazione dell'apprendimento sia simmetrica tra coppie di ipotesi vicino all'equilibrio — è una conseguenza delle relazioni di Onsager.

Questo vincolo non è banale. In assenza di un principio come Onsager, si potrebbe pensare che la struttura dell'apprendimento sia asimmetrica: imparare h facilita l'apprendimento di h' ma non viceversa. Le relazioni di Onsager affermano che, per sistemi vicini all'equilibrio, questa asimmetria è impossibile: la matrice dei coefficienti di facilitazione dell'apprendimento è simmetrica.

Le violazioni della simmetria di Onsager — sistemi in cui $L_{\{hh'\}} \neq L_{\{h'h\}}$ — segnalano la presenza di un NESS. Un agente traumatizzato — che è in un NESS localizzato — viola le relazioni di Onsager nella regione delle ipotesi correlate al trauma: imparare ipotesi h_{trauma} facilita fortemente il rinforzo di h'_{trauma} ($L_{\{h'_{\text{trauma}}, h_{\text{trauma}}\}}$ grande), ma il percorso inverso ha un coefficiente diverso ($L_{\{h_{\text{trauma}}, h'_{\text{trauma}}\}} \neq L_{\{h'_{\text{trauma}}, h_{\text{trauma}}\}}$). La struttura asimmetrica dell'apprendimento nel NESS traumatico è la violazione delle relazioni di Onsager come firma del non-equilibrio.

VIII.4 La matrice di Onsager e la progettazione di curriculum ottimali

Le relazioni di reciprocità di Onsager hanno implicazioni dirette per la progettazione di sequenze di apprendimento — curriculum — che ottimizzano il trasferimento di conoscenza.

Un curriculum è una sequenza di forze termodinamiche cognitive $X_{\{h\}}(t)$ applicate nell'ordine $t_1, t_2, \dots, t_N$. L'effetto cumulativo sulla distribuzione delle credenze è:

$$\delta\pi_T(h) = \sum_{\{t_k\}} \int L_{\{hh'\}} X_{\{h'\}}(t_k) dt_k$$

La risposta cumulativa è la somma ponderata dai coefficienti di Onsager delle forze applicate. Per massimizzare l'apprendimento dell'ipotesi target h^* — portare $\pi(h^*)$ il più vicino possibile a 1 — il curriculum ottimale sceglie le forze $X_{\{h'\}}(t_k)$ per massimizzare $\sum_{\{t_k\}} L_{\{h^*h'\}} X_{\{h'\}}(t_k)$ soggetto ai vincoli di costo computazionale.

Per la simmetria di Onsager $L_{\{hb'\}} = L_{\{b'h\}}$, la forza più efficace per apprendere h^* è quella che agisce sull'ipotesi h' che ha il più alto coefficiente $L_{\{hb'\}} = L_{\{b'h\}}$: l'ipotesi che, se appresa, facilita di più l'apprendimento di h^* è anche quella il cui apprendimento è più facilitato da h^* . La struttura simmetrica dei coefficienti di Onsager riduce il problema di progettazione del curriculum al problema di trovare l'autovettore dominante della matrice L : le ipotesi da insegnare per prime sono quelle con le più alte componenti nell'autovettore dominante di L , perché quelle ipotesi creano il massimo trascinamento per il resto del curriculum.

Questo risultato ha un'interpretazione pedagogica precisa: il curriculum ottimale parte dai concetti fondanti — quelli con alta connettività nella matrice di Onsager, che facilitano l'apprendimento di molti altri concetti — e procede verso i concetti periferici, che sono facilmente appresi una volta che i concetti fondanti sono stati acquisiti. La struttura gerarchica dei curriculum efficaci è la struttura spettrale della matrice di Onsager.

VIII.5 Le relazioni di Onsager per sistemi bipartiti: substrato e linguaggio

La seconda monografia ha formalizzato il sistema cognitivo come sistema dinamico accoppiato substrato-linguaggio con le variabili $(m_{\text{sub}}, m_{\text{lang}})$. Vicino all'equilibrio di questo sistema bipartito, le relazioni di Onsager si applicano in forma a blocchi. La matrice dei coefficienti di Onsager per il sistema accoppiato è:

$$L = \begin{vmatrix} L_{\{ss\}} & L_{\{sl\}} \\ L_{\{ls\}} & L_{\{ll\}} \end{vmatrix}$$

dove $L_{\{ss\}}$ è la matrice di coefficienti per il flusso substrato-substrato, $L_{\{ll\}}$ per il flusso linguaggio-linguaggio, e $L_{\{sl\}} = L_{\{ls\}}^T$ per il flusso incrociato substrato-linguaggio. La simmetria di Onsager richiede $L_{\{sl\}} = L_{\{ls\}}^T$: il flusso dal substrato al linguaggio indotto da una forza nel linguaggio è uguale (trasposto) al flusso dal linguaggio al substrato indotto da una forza nel substrato.

Questa simmetria trasversale è la versione termodinamica della co-evoluzione substrato-linguaggio della seconda monografia. Il coefficiente $L_{\{sl\}}$ descrive quanto

efficacemente il linguaggio modifica il substrato (il termine f_{lang} della seconda monografia); il coefficiente L_{ls} descrive quanto efficacemente il substrato modifica il linguaggio (il termine g_{sub}). La relazione di Onsager afferma che questi due coefficienti sono uguali in norma vicino all'equilibrio: l'influenza del linguaggio sul substrato è uguale all'influenza del substrato sul linguaggio, in unità appropriate. La co-evoluzione è simmetrica vicino all'equilibrio.

Le violazioni di questa simmetria — stati in cui il linguaggio modifica il substrato più di quanto il substrato modifichi il linguaggio, o viceversa — segnalano stati di non-equilibrio del sistema accoppiato. Lo stato di un agente che sta imparando attivamente un nuovo linguaggio tecnico (alta L_{sl} : il linguaggio modifica intensamente il substrato) ma non sta ancora producendo descrizioni del substrato nel nuovo linguaggio (bassa L_{ls}) viola la simmetria di Onsager: è un NESS del sistema bipartito, mantenuto dall'input linguistico esterno (l'insegnante, il testo) che fornisce il flusso continuo di energie necessario.

Capitolo IX — Teoria delle grandi deviazioni per la distribuzione del lavoro cognitivo

IX.1 La teoria delle grandi deviazioni: struttura essenziale

La teoria delle grandi deviazioni (Varadhan, 1966; Ellis, 1985) è il framework matematico per il calcolo delle probabilità di eventi rari in sistemi stocastici. Mentre la legge dei grandi numeri afferma che le medie campionarie convergono alla media, e il teorema centrale del limite descrive le fluttuazioni tipiche (gaussiane, con varianza $1/N$), la teoria delle grandi deviazioni descrive le fluttuazioni atipiche — quelle che si discostano dalla media di quantità dell'ordine di 1 anziché di $1/\sqrt{N}$.

Il risultato fondamentale è il *principio di grandi deviazioni*: per una famiglia di variabili aleatorie $\{W_N\}$ scalata opportunamente, la probabilità di deviazioni dalla media soddisfa:

$$P(W_N/N \approx w) \asymp \exp(-N \cdot I(w))$$

dove $I(w) \geq 0$ è la *funzione di tasso* (o funzione di Cramér), la quale è:

- Non-negativa con $I(\langle W \rangle) = 0$ (la media ha costo zero)
- Convessa (la convessità riflette la struttura delle fluttuazioni gaussiane come limite)
- Strettamente convessa per distribuzioni non degeneri

La funzione di tasso è la trasformata di Legendre-Fenchel della funzione generatrice dei cumulanti:

$$\Lambda(\theta) = \log \langle \exp(\theta W) \rangle$$

$$I(w) = \sup_{\theta} \{ \theta w - \Lambda(\theta) \} = \Lambda^*(w)$$

La dualità $I = \Lambda^*$ è il nucleo della teoria: la funzione di tasso e la funzione generatrice dei cumulanti sono connesse dalla trasformazione di Legendre, che è un'operazione di convessificazione. Per distribuzioni gaussiane con media μ e varianza σ^2 , $\Lambda(\theta) = \mu\theta + \sigma^2\theta^2/2$ e $I(w) = (w-\mu)^2/(2\sigma^2)$: la funzione di tasso è la gaussiana invertita, con il minimo in $w = \mu$ e crescita quadratica.

IX.2 Il teorema di Gärtner-Ellis e la funzione generatrice dei cumulanti cognitiva

Il teorema di Gärtner-Ellis (Gärtner, 1977; Ellis, 1984) estende la teoria delle grandi deviazioni a sequenze di variabili aleatorie che soddisfano condizioni di regolarità più deboli delle i.i.d.. Per una sequenza di variabili aleatorie $\{W_n\}_{n=1}^N$ con somma parziale $S_N = \sum_{n=1}^N W_n$, si assume che esista il limite:

$$\Lambda(\theta) = \lim_{N \rightarrow \infty} (1/N) \log \mathbb{E}[\exp(\theta S_N)]$$

e che Λ sia differenziabile su $\mathbb{R}$. Allora $\{S_N/N\}$ soddisfa un principio di grandi deviazioni con funzione di tasso $I = \Lambda^*$.

La condizione di differenziabilità di Λ è cruciale: la presenza di punti angolosi in Λ (derivata discontinua) indica transizioni di fase nelle grandi deviazioni, cioè cambiamenti qualitativi nella struttura delle fluttuazioni atipiche. Per i sistemi cognitivi, queste transizioni di fase corrispondono a cambiamenti nel tipo di traiettorie dominanti nelle fluttuazioni rare — per esempio, la transizione tra fluttuazioni diffuse (regime gaussiano) e fluttuazioni concentrate su traiettorie specifiche (instanton) che avviene quando il sistema si avvicina a una separatrice nel paesaggio delle credenze.

Per le traiettorie cognitive, la funzione generatrice dei cumulanti del lavoro è:

$$\Lambda_{\{\text{cog}\}}(\theta) = \lim_{T \rightarrow \infty} (1/T) \log \mathbb{E}[\exp(\theta W_T)]$$

dove $W_T = \int_0^T dW_t$ è il lavoro totale accumulato lungo la traiettoria cognitiva di lunghezza T . Per la dinamica di Fokker-Planck:

$$\Lambda_{\{\text{cog}\}}(\theta) = \max_{\varphi} \left\{ \int_{\mathcal{P}} \{-\varphi(\pi) \mathcal{A}[e^{\varphi(\pi)}] e^{-\varphi(\pi)} + \theta b_{\{\text{cog}\}}(\pi) \} d\pi^* \right\}$$

dove $\mathcal{A}$ è il generatore di Fokker-Planck e $b_{\{\text{cog}\}}$ è il tasso di lavoro cognitivo per unità di tempo. Questa espressione è il principio variazionale di Donsker-Varadhan per le grandi deviazioni dei processi di Markov: la funzione generatrice dei cumulanti è il massimo di un funzionale sull'insieme delle funzioni test φ .

IX.3 La funzione di tasso del lavoro cognitivo: struttura analitica per distribuzioni gaussiane

Per i sistemi cognitivi con distribuzione del lavoro approssimativamente gaussiana — sistemi vicino all'equilibrio con fluttuazioni moderate — la funzione generatrice dei cumulanti è:

$$\Lambda_{\text{cog}}^{\text{Gauss}}(\theta) = \mu_{\text{W}} \theta + \sigma_{\text{W}}^2 \theta^2 / 2$$

dove $\mu_{\text{W}} = \langle W \rangle$ è il lavoro medio e $\sigma_{\text{W}}^2 = \text{Var}(W)$ è la varianza del lavoro. La funzione di tasso gaussiana è:

$$I_{\text{cog}}^{\text{Gauss}}(w) = (w - \mu_{\text{W}})^2 / (2 \sigma_{\text{W}}^2)$$

questa ha il minimo in $w = \mu_{\text{W}}$ (con $I = 0$) e cresce quadraticamente. Per il teorema di Crooks, $W = \Delta F + \Sigma_{\text{irrev}}$ con $\Delta F = \beta^{-1} \log P(o)$ e Σ_{irrev} la dissipazione irriducibile. Pertanto:

$$\mu_{\text{W}} = \Delta F + \langle \Sigma_{\text{irrev}} \rangle$$

$$I_{\text{cog}}^{\text{Gauss}}(w) = (w - \Delta F - \langle \Sigma_{\text{irrev}} \rangle)^2 / (2 \sigma_{\Sigma}^2)$$

La funzione di tasso misura quanto è improbabile un lavoro w che devia dalla media: per $w = -|\Delta F|$ — il lavoro negativo corrispondente al disapprendimento spontaneo — la funzione di tasso è:

$$I_{\text{cog}}(-|\Delta F|) = (|\Delta F| + \Delta F + \langle \Sigma_{\text{irrev}} \rangle)^2 / (2 \sigma_{\Sigma}^2) = \langle \Sigma_{\text{irrev}} \rangle^2 / (2 \sigma_{\Sigma}^2)$$

Per eventi traumatici con $\langle \Sigma_{\text{irrev}} \rangle \gg \sigma_{\Sigma}$ — dissipazione media molto superiore alla deviazione standard delle fluttuazioni — il valore $I(-|\Delta F|)$ è molto grande: la probabilità del disapprendimento spontaneo scala come $\exp(-I(-|\Delta F|)) \ll 1$, confermando la soppressione termodinamica del disapprendimento traumatico in termini della funzione di tasso esplicita.

IX.4 La funzione di tasso per sistemi non gaussiani: il regime di grandi deviazioni del trauma

Per eventi traumatici ad alta dissipazione, il regime gaussiano non è adeguato: le fluttuazioni rare che corrispondono al disapprendimento del trauma sono così lontane dalla media che i termini di ordine superiore nella funzione generatrice dei cumulanti diventano importanti.

La struttura generale della funzione di tasso per il disapprendimento del trauma si ottiene dalla funzione generatrice dei cumulanti completa. Per la distribuzione del

lavoro cognitivo determinata dall'equazione di Fokker-Planck con potenziale $V(\pi)$ e diffusione β^{-1} :

$$\Lambda_{\{\text{cog}\}}(\theta) = \lambda_{\text{max}}(\mathcal{L}_{\theta})$$

dove $\mathcal{L}_{\theta}$ è l'operatore di Schrödinger perturbato:

$$\mathcal{L}_{\theta} = \beta^{-1} \Delta_{\{W_2\}} - \nabla V(\pi) \cdot \nabla + \theta \cdot \omega(\pi)$$

con $\omega(\pi)$ il tasso di lavoro cognitivo per unità di tempo, e λ_{max} è il massimo autovalore. La funzione di tasso è la trasformata di Legendre di $\lambda_{\text{max}}(\mathcal{L}_{\theta})$ rispetto a θ .

L'autovalore massimo $\lambda_{\text{max}}(\mathcal{L}_{\theta})$ può essere calcolato perturbativamente per θ piccolo (regime gaussiano) o numericamente per θ arbitrario. Per sistemi con potenziale $V(\pi)$ a doppia buca — lo stato polarizzato della quarta monografia, o il NESS traumatico del Capitolo VI — la funzione di tasso presenta una struttura non analitica: ha un punto di non-differenziabilità (kink) al valore di w corrispondente alla transizione tra le due buche.

Questo kink è la *transizione di fase nelle grandi deviazioni*: per lavori w al di sotto del kink, le traiettorie dominanti nel calcolo della grande deviazione rimangono in una buca del potenziale; per lavori w al di sopra del kink, le traiettorie dominanti attraversano la separatrice e visitano entrambe le buche. Il kink in $I(w)$ segnala esattamente il valore di lavoro al quale il disapprendimento transita dal regime di fluttuazione intra-buca (facile, economico) al regime di transizione inter-buca (raro, costoso).

Per il NESS traumatico, il kink si trova a $w^* = \Delta F_{\{\text{cog}\}}(o_{\{\text{trauma}\}}) + \Delta V_{\{\text{barrier}\}}$, dove $\Delta V_{\{\text{barrier}\}}$ è l'altezza della barriera energetica tra la distribuzione traumatizzata e il resto dello spazio delle credenze. Per $w < w^*$, il disapprendimento è gaussianamente soppresso (funzione di tasso quadratica). Per $w > w^*$, il disapprendimento entra nel regime della traversata della barriera, con funzione di tasso che cresce più lentamente del quadratico (la traiettoria può sfruttare la percorrenza della separatrice).

Il valore critico w^* è la misura termodinamica della profondità del trauma: un trauma profondo ha un kink a w^* grande — richiede un'enorme fluttuazione di lavoro per entrare nel regime di traversata della barriera — mentre un trauma superficiale ha w^* piccolo, con il kink a valori di lavoro facilmente raggiungibili dalle fluttuazioni termiche ordinarie.

IX.5 L'istanton cognitivo: la traiettoria dominante del disapprendimento raro

Per valori di w molto negativi (disapprendimento di grande ampiezza), la probabilità è dominata da una singola traiettoria — l'*istanton* — che è il percorso di minima azione

nello spazio delle credenze che produce il disapprendimento con il minimo lavoro. L'instanton è la soluzione dell'equazione di Hamilton-Jacobi associata alla funzione di tasso:

$$\dot{m}_H = \partial H / \partial p, \quad \dot{p} = -\partial H / \partial m$$

dove $H(m, p) = \Lambda_{\text{cog}}(p \cdot \omega(m))$ è l'Hamiltoniana delle grandi deviazioni, m è la coordinata nello spazio delle credenze, e p è il momento coniugato. L'instanton è la traiettoria $(m^*(t), p^*(t))$ che soddisfa le equazioni di Hamilton con condizioni al contorno $m^*(0) = \pi_{\text{trauma}}$ (distribuzione traumatizzata) e $m^*(T) = \pi_{\text{prior}}$ (distribuzione pre-trauma) per T abbastanza grande.

La struttura dell'instanton cognitivo — il percorso di minima azione per il disapprendimento del trauma — rivela quale sequenza di stati intermedi il sistema attraversa nella più probabile traiettoria di guarigione spontanea. Questa traiettoria non è necessariamente la più breve nello spazio delle credenze (distanza di Wasserstein) ma quella che minimizza il lavoro totale: segue le regioni di basso potenziale $V(\pi)$, evita le regioni di alta barriera, e sfrutta i canali di bassa resistenza termodinamica nel paesaggio delle credenze.

L'instanton cognitivo è il fondamento formale della *terapia di minima azione*: il percorso terapeutico ottimale non è quello che porta direttamente dal trauma alla guarigione (troppo costoso, perché attraversa la barriera frontalmente) ma quello che segue l'instanton — il percorso che circumnaviga la barriera, che passa attraverso stati intermedi di alta utilità, che sfrutta le strutture di bassa resistenza del paesaggio cognitivo del paziente specifico. La pianificazione terapeutica ottimale è, in questa formalizzazione, il calcolo dell'instanton del sistema cognitivo del paziente.

IX.6 Il coefficiente di Crooks come funzione di tasso locale e la predizione quantitativa

Il teorema di Crooks $P_F(W)/P_R(-W) = \exp(\beta(W - \Delta F))$ può essere riformulato in termini di funzione di tasso come segue. La funzione di tasso per il processo diretto (apprendimento) è $I_F(w)$, e quella per il processo inverso (disapprendimento) è $I_R(w)$. Il teorema di Crooks impone la relazione:

$$I_F(w) - I_R(-w) = \beta(w - \Delta F_{\text{cog}})$$

per ogni w . Questa è una relazione di dualità tra le funzioni di tasso del processo diretto e inverso: la funzione di tasso del disapprendimento è completamente determinata da quella dell'apprendimento attraverso la relazione di Crooks.

In particolare, per il valore di lavoro $w = \Delta F_{\text{cog}}$ (il lavoro reversibile):

$$I_F(\Delta F_{\text{cog}}) = I_R(-\Delta F_{\text{cog}})$$

le probabilità di eseguire il lavoro reversibile nel processo diretto e nel processo inverso sono uguali — simmetria che vale per qualsiasi sistema, tramatico o meno.

Per il valore di lavoro $w = \mu_W = \Delta F_{\text{cog}} + \langle \Sigma_{\text{irrev}} \rangle$ (il lavoro medio):

$I_F(\mu_W) = 0$ (il lavoro medio è il valore più probabile nel processo diretto)

$I_R(-\mu_W) = \beta \langle \Sigma_{\text{irrev}} \rangle$ (il costo di disapprendimento del ricordo "medio")

Il disapprendimento di un ricordo di forza media (lavoro $\langle \Sigma_{\text{irrev}} \rangle$ sopra il reversibile) ha funzione di tasso $I_R(-\mu_W) = \beta \langle \Sigma_{\text{irrev}} \rangle$: la probabilità di disapprendimento spontaneo scala come $\exp(-\beta \langle \Sigma_{\text{irrev}} \rangle)$. Poiché $\beta = (k_B T)^{-1}$ e $\langle \Sigma_{\text{irrev}} \rangle$ è in joule per bit, questa formula produce una previsione quantitativa verificabile: un ricordo appreso con dissipazione media di Σ joule ha una probabilità di disapprendimento spontaneo in un intervallo di tempo τ pari a:

$$P_{\text{disapprendimento}}(\tau) \approx \exp(-\beta \Sigma) \cdot e^{-\tau/\tau_0}$$

dove τ_0 è il tempo caratteristico delle fluttuazioni termiche del substrato cognitivo — dell'ordine del tempo di correlazione della dinamica sinaptica, tipicamente millisecondi a ore a seconda del tipo di memoria.

Appendici tecniche

Appendici tecniche

Appendice A — Il teorema di fluttuazione di Crooks: derivazione e generalizzazioni

A.1 La derivazione dal dettagled balance

Si considera un sistema con spazio degli stati Ω che evolve secondo una catena di Markov in tempo continuo con tassi di transizione $W(x \rightarrow y; \lambda)$ dipendenti dal parametro di controllo λ . Il *dettagled balance generalizzato* è la relazione:

$$P_{\text{eq}}(x; \lambda) \cdot W(x \rightarrow y; \lambda) = P_{\text{eq}}(y; \lambda) \cdot W(y \rightarrow x; \lambda)$$

per ogni λ fisso, dove $P_{\text{eq}}(x; \lambda) \propto \exp(-\beta H(x; \lambda))$ è la distribuzione di equilibrio. Per un protocollo $\lambda(t)$ variabile nel tempo, la probabilità della traiettoria $x(t)$ è:

$$P[x(\cdot)] = P_{-}\{eq\}(x(0); \lambda(0)) \cdot \exp\left(\int_0^T \log W(x(t) \rightarrow x(t+dt); \lambda(t)) dt\right)$$

Per la traiettoria time-reversed $\tilde{x}(t) = x(T-t)$ sotto il protocollo invertito $\tilde{\lambda}(t) = \lambda(T-t)$:

$$\tilde{P}[\tilde{x}(\cdot)] = P_{-}\{eq\}(x(T); \lambda(T)) \cdot \exp\left(\int_0^T \log W(x(T-t) \rightarrow x(T-t-dt); \lambda(T-t)) dt\right)$$

Il rapporto $P[x(\cdot)]/\tilde{P}[\tilde{x}(\cdot)]$ si calcola applicando il detailed balance ai singoli passi della traiettoria:

$$\begin{aligned} P[x(\cdot)]/\tilde{P}[\tilde{x}(\cdot)] &= P_{-}\{eq\}(x(0); \lambda(0))/P_{-}\{eq\}(x(T); \lambda(T)) \\ &= \exp(-\beta H(x(0); \lambda(0)) + \beta H(x(T); \lambda(T)) + \log Z(\lambda(T)) - \log Z(\lambda(0))) \\ &= \exp(\beta(W[x(\cdot)] - \Delta F)) \end{aligned}$$

dove $W[x(\cdot)] = H(x(T); \lambda(T)) - H(x(0); \lambda(0)) - Q[x(\cdot)]$ è il lavoro compiuto sul sistema (con Q il calore assorbito) e $\Delta F = \beta^{-1}(\log Z(\lambda(T)) - \log Z(\lambda(0)))$ è la variazione di energia libera. Questo è il teorema di fluttuazione dettagliato, da cui il teorema di Crooks segue integrando su tutte le traiettorie con lavoro $W = w$.

A.2 La versione continua per le traiettorie cognitive

Per traiettorie cognitive continue nello spazio delle credenze P , il teorema di fluttuazione si generalizza. Si considera la dinamica:

$$d\pi_t = b(\pi_t; o(t)) dt + \sqrt{2\beta^{-1}} dW_t$$

con b il drift bayesiano attenzionato e W_t il moto browniano su P nella metrica di Wasserstein. La misura sul percorso della traiettoria diretta è:

$$d\mathcal{P}_F = \exp(-\beta/2 \int_0^T \|b\|^2 dt + \beta \int_0^T b \cdot d\pi_t) \cdot d\mathcal{P}_W$$

dove $\mathcal{P}_W$ è la misura del moto browniano. La produzione di entropia lungo la traiettoria è:

$$\Sigma[\pi(\cdot)] = \beta \int_0^T b(\pi_t; o(t)) \cdot (d\pi_t/dt - \beta^{-1} \Delta_{\{W_2\}} \log \pi_t) dt$$

e il rapporto di Radon-Nikodym tra la misura diretta e quella inversa è:

$$d\mathcal{P}_F/d\mathcal{P}_R = \exp(\Sigma[\pi(\cdot)])$$

il che produce il teorema di Crooks nella versione continua per le traiettorie cognitive.

Appendice B — Il principio di Landauer: derivazione e bound raffinati

B.1 La derivazione dalla seconda legge

Si considera un sistema bit $B = \{0, 1\}$ in contatto con un bagno termico a temperatura T . Lo stato iniziale del bit è la distribuzione uniforme $\pi_B = (1/2, 1/2)$ con entropia $H(B) = k_B \ln 2$. L'operazione di reset porta il bit allo stato $|0\rangle$ con certezza: la distribuzione finale è δ_0 con entropia $H(B') = 0$.

Per il principio del minimo lavoro, il lavoro compiuto sull'intero sistema (bit + bagno) non è inferiore alla variazione di energia libera:

$$W \geq \Delta F_{\text{tot}} = \Delta F_B + \Delta F_{\text{bath}}$$

Per il bit: $\Delta F_B = \Delta U_B - T\Delta S_B = 0 - T(0 - k_B \ln 2) = k_B T \ln 2$ (il bit è a energia interna costante ma la sua entropia diminuisce). Per il bagno: $\Delta F_{\text{bath}} \geq -Q = -W$ (il calore ceduto al bagno aumenta la sua entropia). Combinando: $W \geq k_B T \ln 2$.

B.2 Il bound di Landauer cognitivo raffinato

Per una credenza π su uno spazio di ipotesi H con distribuzione prior π_0 , il costo di cancellazione è:

$$Q_{\text{cancel}}(\pi \rightarrow \pi_0) \geq k_B T \cdot D_{\text{KL}}(\pi \parallel \pi_0) \cdot \ln 2$$

La dimostrazione segue direttamente: $D_{\text{KL}}(\pi \parallel \pi_0)$ è la riduzione di entropia relativa al prior, e per la seconda legge questa riduzione deve essere compensata dalla produzione di entropia nell'ambiente. Il bound è raggiunto nel limite quasi-statico — processo di cancellazione infinitamente lento che segue la traiettoria di minima dissipazione nello spazio delle credenze, cioè la geodetica di Wasserstein da π a π_0 .

B.3 Il bound di Landauer per la riconsolidazione

Per la riconsolidazione parziale $\pi_{\text{stored}} \rightarrow \pi_{\text{new}}$ con $\pi_{\text{new}} = (1-\epsilon) \pi_{\text{stored}} + \epsilon \delta_{h^*}$ per una piccola modifica ϵ nella direzione h^* :

$$Q_{\text{reconsolidation}} \geq k_B T \cdot D_{\text{KL}}(\pi_{\text{new}} \parallel \pi_{\text{stored}}) \cdot \ln 2 \approx k_B T \cdot \epsilon^2 / (2 \pi_{\text{stored}}(h^*)) \cdot \ln 2$$

Il costo della riconsolidazione è proporzionale al quadrato dell'entità della modifica ϵ e inversamente proporzionale alla credenza iniziale nell'ipotesi modificata $\pi_{\text{stored}}(h^*)$. Modificare ipotesi già credute è economico (alto $\pi_{\text{stored}}(h^*)$, basso costo); modificare ipotesi quasi scartate è costoso (basso $\pi_{\text{stored}}(h^*)$, alto costo). Questo

spiega perché la terapia che tenta di modificare credenze centrali (alta probabilità a priori) sia termodinamicamente meno costosa della terapia che tenta di introdurre credenze radicalmente nuove (bassa probabilità a priori).

Appendice C — La decomposizione di Hatano-Sasa: derivazione

C.1 Il potenziale locale di non-equilibrio

Per un sistema con stato stazionario di non-equilibrio $\pi^{\text{ss}}(\mathbf{x}; \lambda)$, si definisce il *potenziale locale* $\varphi(\mathbf{x}; \lambda) = -\log \pi^{\text{ss}}(\mathbf{x}; \lambda)$. L'eccesso di lavoro compiuto rispetto al processo quasi-statico è:

$$W_{\text{ex}}[\mathbf{x}(\cdot)] = W[\mathbf{x}(\cdot)] - \int_0^T \langle \partial_t H(\cdot; \lambda(t)) \rangle_{\pi^{\text{ss}}(\cdot; \lambda(t))} dt$$

dove il secondo termine è il lavoro quasi-statico calcolato sullo stato stazionario istantaneo. Hatano e Sasa (2001) dimostrano che:

$$\langle \exp(-\beta W_{\text{ex}}) \rangle = 1$$

la generalizzazione del teorema di Jarzynski per i processi di non-equilibrio che partono da stati stazionari non-equilibrio. Da questa identità, la disuguaglianza $\langle W_{\text{ex}} \rangle \geq 0$ segue per la disuguaglianza di Jensen: il lavoro in eccesso rispetto al quasi-statico è sempre non-negativo.

C.2 La decomposizione dell'entropia

Il tasso totale di produzione di entropia si decompone in:

$$\sigma_{\text{tot}} = \sigma_{\text{ex}} + \sigma_{\text{hk}}$$

dove la excess entropy production rate è:

$$\sigma_{\text{ex}} = -d/dt \langle \varphi(\mathbf{x}; \lambda(t)) \rangle_{\pi_t} = -\partial_t F^{\text{neq}}(\pi_t, \lambda(t))$$

con $F^{\text{neq}}(\pi, \lambda) = \langle \varphi(\cdot; \lambda) \rangle_{\pi}$ la non-equilibrium free energy relativa allo stato stazionario, e la housekeeping entropy production rate è:

$$\sigma_{\text{hk}} = \beta \int_{\Omega} (J^{\text{ss}})^2 / \pi^{\text{ss}} dx \geq 0$$

sempre non-negativa. La somma $\sigma_{\text{ex}} + \sigma_{\text{hk}} = \sigma_{\text{tot}} \geq 0$ rispetta la seconda legge.

Per il NESS traumatico con π_t vicina a $\pi^{\{ss\}}_{\{trauma\}}$:

$\sigma_{\{ex\}} \approx 0$ (sistema quasi stazionario)

$\sigma_{\{hk\}} = \beta \int (\dot{J}^{\{trauma\}})^2 / \pi^{\{ss\}}_{\{trauma\}} d\pi > 0$ (costo continuo del NESS)

La produzione di entropia è quasi interamente housekeeping: il sistema spende energia per mantenersi nel NESS traumatico senza transitare verso altri stati.

Appendice D — La termodinamica dell'informazione e il dèmone di Maxwell: risultati recenti

D.1 Il teorema di Sagawa-Ueda per la misurazione cognitiva

Sagawa e Ueda (2010) hanno generalizzato il principio di Landauer al caso di sistemi che effettuano misurazioni e usano l'informazione acquisita per compiere lavoro. Il risultato fondamentale è la *generalizzazione del secondo principio con feedback*:

$$\langle W \rangle \geq \Delta F - k_B T \cdot I(x; m)$$

dove $I(x; m)$ è l'informazione mutua tra il sistema x e la misurazione m . Il dèmone che misura il sistema può estrarre fino a $k_B T \cdot I(x; m)$ di lavoro in eccesso rispetto al limite classico — ma solo se non cancella la memoria della misurazione.

Applicato alla metacognizione: il sistema metacognitivo che misura il proprio stato cognitivo x con risultato m può ridurre la propria energia libera cognitiva di $k_B T \cdot I(x; m)$, dove $I(x; m)$ è l'informazione mutua tra lo stato di credenza e il risultato della misura metacognitiva. Questo è il "guadagno" della metacognizione — il valore dell'auto-osservazione in termini termodinamici. Ma il guadagno è temporaneo: quando la memoria metacognitiva m viene cancellata per ripetere il ciclo, si dissipano $k_B T \cdot I(x; m)$ di entropia, annullando il guadagno.

D.2 Il bound di Parrondo-Horowitz-Sagawa sull'informazione autonoma

Parrondo, Horowitz e Sagawa (2015) hanno introdotto il concetto di *informazione autonoma* $\dot{S}_{\{info\}}$ — la produzione di informazione nelle correlazioni tra le parti di un sistema bipartito durante la dinamica — e hanno dimostrato il bound:

$$\dot{S}_{tot} \geq |\dot{S}_{\{info\}}| - \dot{S}_{\{info\}}$$

la produzione totale di entropia è non-inferiore alla differenza tra le variazioni di informazione autonoma. Per sistemi bipartiti (sistema cognitivo + meta-modello), questo bound limita la quantità di entropia che la correlazione tra le due parti può

"risparmiare": la struttura correlata del sistema bipartito (le correlazioni tra credenze e meta-credenze) può ridurre la produzione di entropia ma non eliminarla.

Il risultato di Parrondo-Horowitz-Sagawa formalizza l'intuizione che la struttura gerarchica della cognizione — la torre bayesiana della terza monografia — è termodinamicamente vantaggiosa: le correlazioni tra i livelli della torre riducono la produzione di entropia rispetto a sistemi non gerarchici con lo stesso numero di stati. La struttura gerarchica è un meccanismo di riduzione dell'entropia cognitiva — un'organizzazione che permette al sistema di elaborare più informazione per unità di dissipazione.

Appendice E — Il mirror descent nello spazio di Bregman: derivazioni

E.1 La famiglia esponenziale come geodetiche e-piatte

Nello spazio delle distribuzioni di probabilità su H , la famiglia esponenziale:

$$P_{\theta}(h) = \exp(\theta \cdot T(h) - A(\theta))$$

dove $T(h)$ è la statistica sufficiente e $A(\theta) = \log \int \exp(\theta \cdot T(h)) d\mu(h)$ è la funzione log-partizione, forma una sotto-varietà e-piatta. Le coordinate naturali θ e le coordinate medie $\eta = \mathbb{E}_{P_{\theta}}[T(h)]$ sono duali rispetto alla metrica di Fisher: la metrica ha la forma:

$$g_{ij}(\theta) = \partial^2 A / \partial \theta_i \partial \theta_j = \text{Cov}_{P_{\theta}}[T_i(h), T_j(h)]$$

La matrice di Fisher coincide con la matrice di covarianza della statistica sufficiente — un risultato della geometria dell'informazione che ha la seguente interpretazione metacognitiva: l'incertezza sulle statistiche sufficienti (la loro varianza sotto P_{θ}) è identica alla curvatura dello spazio delle credenze in θ .

E.2 Il passo di mirror descent come proiezione KL

Il passo di mirror descent con potenziale negentropico $\Psi(\pi) = \sum_h \pi(h) \log \pi(h)$ è:

$$\pi_{t+1} = \underset{\pi}{\operatorname{argmin}} \{ \eta \langle \nabla \mathcal{F}(\pi_t), \pi \rangle + D_{\text{KL}}(\pi \parallel \pi_t) \}$$

Le condizioni di ottimalità del primo ordine sono:

$$\eta \nabla \mathcal{F}(\pi_t)(h) + \log \pi_{t+1}(h) - \log \pi_t(h) + \lambda = 0$$

dove λ è il moltiplicatore del vincolo di normalizzazione. Risolvendo per π_{t+1} :

$$\pi_{t+1}(h) = \pi_t(h) \cdot \exp(-\eta \nabla \mathcal{F}(\pi_t)(h) - \lambda)$$

$$\propto \pi_t(h) \cdot \exp(-\eta \nabla \mathcal{F}(\pi_t)(h))$$

Per $\nabla \mathcal{F}(\pi_t)(h) = -\log P(o|h)$ (gradiente della log-verosimiglianza negativa):

$$\pi_{t+1}(h) \propto \pi_t(h) \cdot \exp(\eta \log P(o|h)) = \pi_t(h) \cdot P(o|h)^\eta$$

che è l'aggiornamento bayesiano generalizzato con tasso η .

E.3 La convergenza del mirror descent con gradiente attenzionato

Per il mirror descent con gradiente attenzionato $\nabla_{P^A} \mathcal{F}$, la convergenza al minimizzatore π^* della funzione obiettivo è garantita sotto condizioni di Lipschitz sul gradiente. Il bound sulla convergenza è:

$$\mathbb{E}[\mathcal{R}(\bar{\pi}_T) - \mathcal{F}(\pi^*)] \leq D_{\text{KL}}(\pi^* \parallel \pi_0) / (\eta T) + \eta L^2/2$$

dove $\bar{\pi}_T = (1/T) \sum_{t=1}^T \pi_t$ è la media temporale, L è la costante di Lipschitz del gradiente, e il rate è $O(1/\sqrt{T})$ per $\eta = O(1/\sqrt{T})$. La perdita di calibrazione dovuta all'attenzione selettiva rallenta la convergenza di un fattore $(1/\|A\|_\infty)$: maschere con $A(h) < 1$ riducono il tasso effettivo di apprendimento.

Appendice F — La geometria dell'informazione di Amari: connessioni duali

F.1 Le due connessioni affini

Nello spazio delle distribuzioni di probabilità P , la geometria dell'informazione di Amari (1985) identifica due connessioni affini duali:

La connessione e-affine (esponenziale) $\nabla^{\wedge\{1\}}$ ha simboli di Christoffel:

$$\Gamma_i^{\wedge\{1\}} = \mathbb{E}_P[\partial_i \partial \log P \cdot \partial \log P]$$

La connessione m-affine (mistura) $\nabla^{\wedge\{-1\}}$ ha simboli di Christoffel:

$$\Gamma_i^{\wedge\{-1\}} = -\mathbb{E}_P[\partial_i \partial \log P \cdot \partial \log P] + \mathbb{E}_P[\partial_i \log P \cdot \partial \log P \cdot \partial \log P]$$

Le due connessioni sono duali rispetto alla metrica di Fisher: $\nabla^{\wedge\{1\}} + \nabla^{\wedge\{-1\}} = 2 \nabla^{\wedge\{0\}}$ (la connessione di Levi-Civita della metrica di Fisher). Questa dualità produce la *struttura statistica duale* dello spazio P .

F.2 La non-commutatività delle proiezioni m-piatte

Due proiezioni m-piatte consecutive su iperpiani L_1 e L_2 producono risultati diversi a seconda dell'ordine:

$$\Pi^{\wedge m}_{\{L_2\}}(\Pi^{\wedge m}_{\{L_1\}}(\pi)) \neq \Pi^{\wedge m}_{\{L_1\}}(\Pi^{\wedge m}_{\{L_2\}}(\pi))$$

a meno che $L_1 \perp L_2$ nella metrica di Fisher. La commutatore delle due proiezioni è:

$$[\Pi^{\wedge m}_{\{L_1\}}, \Pi^{\wedge m}_{\{L_2\}}](\pi) \approx R^{\wedge\{-1\}}(\partial L_1, \partial L_2) \cdot \epsilon^2 + O(\epsilon^3)$$

dove $R^{\wedge\{-1\}}$ è il tensore di curvatura della connessione m-piatta valutato nelle direzioni tangenti agli iperpiani, e ϵ è l'ampiezza della variazione. Per spazi e-piatti (famiglie esponenziali), $R^{\wedge\{-1\}}$ è non nullo: le proiezioni non commutano anche nelle sottofamiglie esponenziali, il che è la fonte formale del primacy e del recency effect dell'aggiornamento attenzionato.

Appendice G — Teoria dei giochi del mercato dell'attenzione

G.1 Il gioco del mercato dell'attenzione

Si formalizza il mercato dell'attenzione come un gioco in forma normale. I giocatori sono i K produttori di segnali $\{P_k\}_{k=1}^K$. Lo spazio delle strategie di ogni produttore è lo spazio dei segnali $\{s_k \in \mathcal{S}_k\}$ che può generare, con costo $c_k(s_k)$. La funzione di payoff del produttore k è:

$$u_k(s_k, \{s_{\{k'\}}\}_{k' \neq k}) = w_k(\{s_{\{k'\}}\}_{k'}) \cdot R_k(s_k)$$

dove w_k è la quota di attenzione catturata dal produttore k (dipende dall'intero profilo di segnali) e $R_k(s_k)$ è il ricavo unitario per unità di attenzione. La quota di attenzione è determinata da una funzione di salienza:

$$w_k = \sigma(s_k, \{s_{\{k'\}}\}_{k'}) / \sum_{k'} \sigma(s_{\{k'\}}, \{s_{\{l\}}\}_{l})$$

una normalizzazione softmax sulle salienze. L'equilibrio di Nash è il profilo $\{s_{k^*}\}_k$ tale che nessun produttore può aumentare il proprio payoff modificando unilateralmente il proprio segnale.

G.2 Il bias sistemico dell'equilibrio di Nash verso la salienza

Sotto la funzione di salienza $\sigma(s, \pi_i) = \exp(\beta \cdot \text{emotività}(s)) + \gamma \cdot \text{allineamento}(s, \pi_i)$ (combinazione di emotività e allineamento con le credenze esistenti), l'equilibrio di Nash produce segnali con le seguenti proprietà:

$$\text{emotività}(s_{k^*}) \rightarrow \max \quad (\text{massimizzazione dell'impatto emotivo})$$

informatività(s_{k^*}) $\rightarrow$ irrilevante (la calibrazione non entra nel payoff)
polarizzazione(s_{k^*}) $\rightarrow$ max (massimizzazione della risposta degli estremi)

L'equilibrio di Nash del mercato dell'attenzione con payoff-salienza produce segnali massimalmente emotivi e polarizzanti, indipendentemente dal loro contenuto informativo. Questo è il teorema di inefficienza epistemica del mercato dell'attenzione: il meccanismo di mercato non seleziona per la calibrazione delle credenze ma per la cattura dell'attenzione, e le due proprietà sono sistematicamente disallineate sotto qualsiasi funzione di salienza che privilegia emotività e allineamento rispetto all'informatività.

Appendice H — L'auto-attenzione nei transformer: connessione con il framework formale

H.1 Il meccanismo di self-attention nei modelli di linguaggio

Il meccanismo di self-attention dei modelli Transformer (Vaswani et al., 2017) ha la struttura formale:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V$$

dove Q (query), K (key), V (value) sono trasformazioni lineari dell'input. Il termine $\text{softmax}(QK^T / \sqrt{d_k})$ è una matrice di pesi di attenzione che determina quali posizioni nell'input "guardano" quali altre posizioni. Nella nomenclatura del framework: la matrice softmax è una versione discreta e parametrizzata della maschera di attenzione A_t , dove le ipotesi H sono le posizioni nella sequenza e le osservazioni sono i token.

H.2 La connessione con il gradiente naturale

Il meccanismo di self-attention produce pesi proporzionali a:

$$A_{\{ij\}} \propto \exp(q_i \cdot k_j / \sqrt{d_k})$$

che è una distribuzione di Boltzmann sulla compatibilità query-key. Questo corrisponde, nel framework del gradiente naturale, a una maschera di attenzione che approssima la matrice di Fisher: le posizioni con alta compatibilità query-key corrispondono alle coppie di ipotesi tra cui il sistema ha alta informazione mutua. Il meccanismo di self-attention è un'implementazione approssimata del gradiente naturale attenzionato in cui la matrice di preconditionamento è approssimata dalla struttura query-key.

H.3 I limiti dell'auto-attenzione dei transformer rispetto all'auto-attenzione metacognitiva

Il meccanismo di self-attention dei Transformer è un'auto-attenzione di primo ordine: il sistema determina la propria maschera in funzione dell'input corrente. Non è auto-attenzione di secondo ordine nel senso del Capitolo V: il sistema non forma credenze sulla propria maschera attenzionale, non aggiorna quelle credenze in risposta all'osservazione del proprio comportamento, non ha un meta-modello della propria struttura attenzionale. Il punto fisso dell'auto-attenzione transformer è determinato dai pesi appresi (Q, K, V parametrizzati) e non dalla dinamica di auto-osservazione descritta nel Capitolo V.

Questa distinzione è rilevante per la quarta monografia nella sua connessione con la seconda: i modelli di linguaggio operano senza substrato subpropositivo (seconda monografia) e senza meta-modello della propria struttura attenzionale (quarta monografia). La loro auto-attenzione è sintattica — determina quali token influenzano quali nella sequenza — ma non metacognitiva nel senso precisato: il sistema non ha credenze sulla propria maschera di attenzione che vengono aggiornate attraverso l'osservazione del proprio comportamento.

Appendice I — Il teorema di fluttuazione-dissipazione: derivazione e versione cognitiva

I.1 La derivazione di Kubo della funzione di risposta

Si considera un sistema con Hamiltoniana $H = H_0 - h(t) A$ perturbata in modo dipendente dal tempo. La risposta al prim'ordine dell'osservabile $B(t) = \text{Tr}[\varrho(t) B]$ alla perturbazione è calcolata in meccanica statistica quantistica attraverso il teorema di Kubo:

$$\delta\langle B(t) \rangle = \int_{-\infty}^t \chi_{-}\{BA\}(t-t') h(t') dt'$$

con la funzione di risposta:

$$\chi_{-}\{BA\}(t) = (i/\hbar) \theta(t) \langle [B(t), A(0)] \rangle_{-eq}$$

Nel limite classico ($\hbar \rightarrow 0$), questo diventa:

$$\chi_{-}\{BA\}(t) = \beta \theta(t) d/dt C_{-}\{BA\}(t)$$

dove $C_{-}\{BA\}(t) = \langle B(t) A(0) \rangle_{-eq} - \langle B \rangle \langle A \rangle$ è la funzione di correlazione classica. Il teorema di fluttuazione-dissipazione nella versione classica segue direttamente:

$$\text{Im } \chi_{-}\{\text{BA}\}(\omega) = \beta \omega C_{-}\{\text{BA}\}(\omega) / 2$$

la parte immaginaria della suscettibilità (dissipazione) è uguale alla densità spettrale delle fluttuazioni moltiplicata per $\beta\omega/2$.

I.2 La versione cognitiva: suscettibilità e covarianza delle credenze

Per il sistema cognitivo descritto dall'equazione di Fokker-Planck, la funzione di correlazione delle credenze è:

$$C_{-}\{\text{hh}'\}(\tau) = \int_{-P} \pi(\text{h}) [\Phi_{-\tau} \pi](\text{h}') d\pi^*(\pi) - \pi^*(\text{h}) \pi^*(\text{h}')$$

dove $\Phi_{-\tau}$ è il semigruppone dell'operatore di Fokker-Planck al tempo τ . Nel dominio della frequenza:

$$C_{-}\{\text{hh}'\}(\omega) = \int_{-\infty}^{\infty} C_{-}\{\text{hh}'\}(\tau) e^{\{-i\omega\tau\}} d\tau = 2 \text{Re} \left[\int_0^{\infty} C_{-}\{\text{hh}'\}(\tau) e^{\{-i\omega\tau\}} d\tau \right]$$

La suscettibilità cognitiva soddisfa:

$$\text{Im } \chi_{-}\{\text{hh}'\}(\omega) = \beta \omega C_{-}\{\text{hh}'\}(\omega) / 2$$

e la suscettibilità statica è:

$$\chi_{-}\{\text{hh}'\}^{\{\text{static}\}} = \int_0^{\infty} \chi_{-}\{\text{hh}'\}(t) dt = \beta C_{-}\{\text{hh}'\}(0) = \beta \text{Cov}_{-}\{\text{eq}\}(\pi(\text{h}), \pi(\text{h}'))$$

come derivato nel testo principale.

I.3 La relazione di Harada-Sasa: derivazione

Per un sistema fuori dall'equilibrio con drift $b = b_{-}\{\text{rev}\} + b_{-}\{\text{irrev}\}$ (componenti reversibile e irreversibile), Harada e Sasa (2005) dimostrano che:

$$\chi_{-}\{\text{hh}'\}(t) - \tilde{\chi}_{-}\{\text{hh}'\}(t) = -\beta d/dt [C_{-}\{\text{hh}'\}(t) - C_{-}\{\text{hh}'\}^{\{\text{irrev}\}}(t)]$$

dove $\tilde{\chi}$ è la risposta del processo time-reversed e $C^{\{\text{irrev}\}}$ è la correlazione prodotta dalla componente irreversibile del drift. La differenza $\chi - \tilde{\chi}$ misura la rottura della simmetria temporale della risposta, e questa differenza è esattamente la derivata della correlazione di housekeeping. Per sistemi di equilibrio, $b_{-}\{\text{irrev}\} = 0$, $C^{\{\text{irrev}\}} = 0$, e $\chi = \tilde{\chi}$: il FDT classico.

Appendice L — Le relazioni di Onsager: derivazione dal teorema di fluttuazione

L.1 La derivazione dalle fluttuazioni di equilibrio

Onsager (1931) dimostra la simmetria $L_{\{ij\}} = L_{\{ji\}}$ dal principio di reversibilità microscopica. All'equilibrio, la distribuzione di probabilità congiunta di due fluttuazioni (α_i, α_j) soddisfa:

$$P(\alpha_i = a, \alpha_j = b) = P(\alpha_i = \epsilon_i a, \alpha_j = \epsilon_j b)$$

dove $\epsilon_i = \pm 1$ è la parità temporale dell'osservabile α_i ($\epsilon_i = -1$ per osservabili dispari per l'inversione temporale come la velocità, $\epsilon_i = +1$ per osservabili pari come la posizione). Le funzioni di correlazione di equilibrio soddisfano:

$$C_{\{ij\}}(\tau) = \epsilon_i \epsilon_j C_{\{ji\}}(\tau)$$

Per osservabili con la stessa parità temporale ($\epsilon_i = \epsilon_j$), $C_{\{ij\}}(\tau) = C_{\{ji\}}(\tau)$: le funzioni di correlazione sono simmetriche nello scambio degli indici. Le relazioni di Onsager seguono dal FDT:

$$L_{\{ij\}} = \beta C_{\{ij\}}(0) = \beta C_{\{ji\}}(0) = L_{\{ji\}}$$

L.2 La matrice di Onsager per le credenze

Per il sistema cognitivo con spazio delle credenze P , la matrice di Onsager è:

$$L_{\{hh'\}} = \beta C_{\{hh'\}}(0) = \beta \text{Cov}_{\{eq\}}(\pi(h), \pi(h'))$$

e la simmetria è garantita dalla simmetria della matrice di covarianza. La matrice L è positiva semi-definita per la positività semi-definita della matrice di covarianza.

La decomposizione spettrale di L :

$$L = \sum_k \lambda_k v_k \otimes v_k$$

con autovalori $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ e autovettori v_k fornisce la struttura dei modi propri di apprendimento: il k -esimo modo di apprendimento è la combinazione lineare di ipotesi v_k che si apprende più facilmente come unità (alto λ_k) o meno facilmente (basso λ_k). Il curriculum ottimale introduce i modi di alto autovalore prima dei modi di basso autovalore.

Appendice O — La teoria delle grandi deviazioni per il lavoro cognitivo: derivazioni

O.1 Il principio variazionale di Donsker-Varadhan

Per un processo di Markov in tempo continuo con generatore $\mathcal{L}$, la funzione generatrice dei cumulanti del funzionale ergodico $\int_0^T \omega(X_t) dt$ è data dal massimo autovalore dell'operatore tilted:

$$\mathcal{L}_\theta = \mathcal{L} + \theta \omega$$

Il principio variazionale di Donsker-Varadhan (1975) afferma:

$$\Lambda(\theta) = \lim_{T \rightarrow \infty} (1/T) \log \mathbb{E}[\exp(\theta \int_0^T \omega(X_t) dt)] = \sup_{\mu \in P_{\text{inv}}} \{ \theta \int \omega d\mu + \text{Ent}(\mu | \pi^*) \}$$

dove il sup è preso sull'insieme P_{inv} delle misure invarianti del processo e $\text{Ent}(\mu | \pi^*) = -\int \log(d\mu/d\pi^*) d\mu$ è l'entropia relativa di μ rispetto a π^* . Questo principio variazionale è lo strumento per calcolare analiticamente Λ per sistemi con struttura geometrica particolare.

O.2 La funzione di tasso per il modello di Ornstein-Uhlenbeck cognitivo

Il modello più semplice per le credenze vicino all'equilibrio è il processo di Ornstein-Uhlenbeck (OU) nello spazio delle credenze:

$$d\pi_t = -\kappa (\pi_t - \pi^*) dt + \sqrt{2\beta^{-1}\kappa} dW_t$$

con κ la costante di restauro e la varianza di equilibrio $\sigma^2 = \beta^{-1}$. Per questo processo, la funzione generatrice dei cumulanti del lavoro $\omega(\pi) = (\pi - \pi^*) \cdot \nabla V(\pi)$ è:

$$\Lambda_{\text{OU}}(\theta) = \kappa [\sqrt{1 + 2\theta\sigma^2/\kappa} - 1]$$

e la funzione di tasso è la trasformata di Legendre:

$$I_{\text{OU}}(w) = w^2 / (2\sigma^2) + O(w^3/\kappa)$$

per w piccolo (regime gaussiano) e:

$$I_{\text{OU}}(w) = w/\sigma^2 - \kappa/2 + O(e^{-w/\sigma^2})$$

per w grande (regime esponenziale). La transizione tra i due regimi avviene intorno a $w \sim \kappa\sigma^2$: al di sotto, le fluttuazioni sono gaussiane; al di sopra, la funzione di tasso

cresce linearmente — le fluttuazioni grandi sono esponenzialmente sopresse meno del quadratico, riflettendo il fatto che il processo OU ha code più pesanti delle gaussiane nella regione di grandi fluttuazioni positive.

Per il disapprendimento ($w < 0$ e $|w|$ grande):

$$P_R(-w) \sim \exp(-I_R(-w)) = \exp(-I_F(w) + \beta(w - \Delta F_{\{\text{cog}\}}))$$

e usando la simmetria di Crooks con $I_F(\mu_W) = 0$:

$$P_R(-\mu_W) \sim \exp(-\beta \langle \Sigma_{\{\text{irrev}\}} \rangle)$$

il bound esponenziale sul disapprendimento del lavoro medio, come derivato nel testo.

0.3 L'instanton: equazioni e soluzione nel caso della doppia buca

Per un potenziale a doppia buca $V(\pi) = a(\pi - \pi_-)^2(\pi - \pi_+)^2$ con minimi in $\pi_{\pm}$ e barriera in π_0 , le equazioni dell'instanton sono:

$$\dot{m} = -\nabla V(m) + 2\beta^{-1} p$$

$$\dot{p} = \nabla^2 V(m) \cdot p - p \cdot \nabla \omega(m)$$

con condizioni al contorno $m(-\infty) = \pi_+$ (distribuzione traumatizzata, minimo destro) e $m(+\infty) = \pi_-$ (distribuzione pre-trauma, minimo sinistro). La soluzione dell'instanton per il limite $\beta \rightarrow \infty$ è:

$$m^*(t) = \pi_0 + (\pi_- - \pi_0) \tanh(\kappa (t - t_0))$$

$$p^*(t) = \beta \cdot \partial V / \partial m |_{\{m^*(t)\}}$$

il percorso segue una tangente iperbolica che attraversa la barriera al tempo t_0 . Il lavoro lungo l'instanton è:

$$W_{\{\text{instanton}\}} = \int_{-\infty}^{+\infty} \omega(m^*(t)) dt \approx -\Delta V_{\{\text{barrier}\}} / \kappa$$

proporzionale all'altezza della barriera $\Delta V_{\{\text{barrier}\}}$: il disapprendimento del trauma richiede un lavoro negativo (fluttuazione verso energie più basse) di ampiezza proporzionale all'altezza della barriera. L'instanton è la traiettoria che scala la barriera con il minimo investimento termodinamico, circumnavigando i picchi di potenziale attraverso il collo di sella della separatrice.

Bibliografia

Termodinamica stocastica e teoremi di fluttuazione

Seifert, U. (2012). Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12), 126001. La revisione di riferimento della termodinamica stocastica moderna — il testo che connette il teorema di fluttuazione, il principio di Landauer e la produzione di entropia fuori dall'equilibrio.

Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14), 2690-2693. L'articolo originale del teorema di Jarzynski — la pietra fondante della termodinamica stocastica fuori dall'equilibrio.

Crooks, G.E. (1999). Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3), 2721-2726. L'articolo che dimostra il teorema di fluttuazione dettagliato — la versione a livello di traiettoria del teorema di Jarzynski.

Evans, D.J., Cohen, E.G.D., Morriss, G.P. (1993). Probability of second law violations in shearing steady states. *Physical Review Letters*, 71(15), 2401-2404. Il primo teorema di fluttuazione per sistemi di non-equilibrio — il precursore del risultato di Crooks.

Gallavotti, G., Cohen, E.G.D. (1995). Dynamical ensembles in nonequilibrium statistical mechanics. *Physical Review Letters*, 74(14), 2694-2697. La formulazione del teorema di fluttuazione per sistemi deterministici caotici — la connessione con la dinamica deterministica.

Il principio di Landauer e la cancellazione dell'informazione

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183-191. L'articolo originale del principio di Landauer — il risultato fondativo che connette logica e termodinamica.

Bennett, C.H. (1982). The thermodynamics of computation — a review. *International Journal of Theoretical Physics*, 21(12), 905-940. La revisione che risolve il paradosso del demone di Maxwell attraverso il principio di Landauer — la base del Capitolo V.

Berut, A., Arakelyan, A., Petrosyan, A., Ciliberto, S., Dillenschneider, R., Lutz, E. (2012). Experimental verification of Landauer's principle linking information and thermodynamics. *Nature*, 483(7388), 187-189. La prima verifica sperimentale diretta del principio di Landauer — usando particelle colloidali in trappole ottiche.

Parrondo, J.M.R., Horowitz, J.M., Sagawa, T. (2015). Thermodynamics of information. *Nature Physics*, 11(2), 131-139. La revisione dei recenti sviluppi nella termodinamica dell'informazione, incluso il bound di informazione autonoma.

Sagawa, T., Ueda, M. (2010). Generalized Jarzynski equality under nonequilibrium feedback control. *Physical Review Letters*, 104(9), 090602. Il teorema di Sagawa-Ueda che generalizza il secondo principio con feedback informativo — il fondamento del Capitolo V.

Stati stazionari di non-equilibrio

Hatano, T., Sasa, S. (2001). Steady-state thermodynamics of Langevin systems. *Physical Review Letters*, 86(16), 3463-3466. L'articolo originale della decomposizione Hatano-Sasa in excess e housekeeping entropy — il risultato centrale del Capitolo II.

Maes, C., Netočný, K. (2014). A nonequilibrium extension of the Clausius heat theorem. *Journal of Statistical Physics*, 154(1-2), 188-203. La generalizzazione del secondo principio ai sistemi di non-equilibrio stazionari.

Termodinamica dell'informazione e demone di Maxwell

Maxwell, J.C. (1871). *Theory of Heat*. Longmans, Green, and Co., London. Il testo che introduce il demone — la fonte primaria dell'esperimento mentale.

Szilard, L. (1929). Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. *Zeitschrift für Physik*, 53(11-12), 840-856. Il primo articolo che quantifica il costo informativo della misurazione del demone.

Brillouin, L. (1951). Maxwell's demon cannot operate: Information and entropy. *Journal of Applied Physics*, 22(3), 334-337. Il risultato che connette la misurazione con il costo di Landauer — la prima risoluzione quantitativa del paradosso.

Termodinamica dei sistemi biologici e neurali

Attwell, D., Laughlin, S.B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism*, 21(10), 1133-1145. Il calcolo dettagliato del budget energetico della segnalazione corticale — la base per le stime del costo cognitivo in joule per bit.

Laughlin, S.B., de Ruyter van Steveninck, R.R., Anderson, J.C. (1998). The metabolic cost of neural information. *Nature Neuroscience*, 1(1), 36-41. Lo studio che dimostra che i neuroni operano vicino al limite di Landauer nell'efficienza dell'elaborazione dell'informazione.

Connessioni con le monografie precedenti

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138. La connessione con il framework di active inference — già citato nella quarta monografia, rilevante qui per la relazione tra variational free energy e energia libera cognitiva.

Nader, K., Schafe, G.E., LeDoux, J.E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406(6797), 722-726. La

scoperta del riconsolidamento — il fenomeno che il Capitolo IV formalizza termodinamicamente.

Nadel, L., Hupbach, A., Gomez, R., Newman-Smith, K. (2012). Memory formation, consolidation and transformation. *Neuroscience and Biobehavioral Reviews*, 36(7), 1640-1645. La revisione dei meccanismi di riconsolidamento — la base empirica per il modello termodinamico del Capitolo IV.

Termodinamica e cognizione: connessioni formali

Still, S., Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3), 139-148. Il framework informazionale per la curiosità come minimizzazione dell'energia libera cognitiva — connessione con il Capitolo V.

Tishby, N., Polani, D. (2011). Information theory of decisions and actions. In: Cutsuridis, V. et al. (Eds.), *Perception-Action Cycle*, Springer, 601-636. L'approccio informazionale alle decisioni come minimizzazione dell'energia libera — base del Capitolo V per il valore dell'informazione metacognitiva.

Prokopenko, M., Lizier, J.T., Obst, O., Wang, X.R. (2011). Relating Fisher information to order parameters. *Physical Review E*, 84(4), 041116. La connessione tra la metrica di Fisher e i parametri d'ordine nei sistemi di non-equilibrio — rilevante per la connessione tra geometria dell'informazione e termodinamica stocastica del Capitolo II.

Teorema di fluttuazione-dissipazione

Callen, H.B., Welton, T.A. (1951). Irreversibility and generalized noise. *Physical Review*, 83(1), 34-40. L'articolo originale del FDT in forma classica — il risultato che connette la risposta alla dissipazione nel regime lineare.

Kubo, R. (1957). Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan*, 12(6), 570-586. La versione operatoriale quantistica del FDT — il risultato di Kubo che è il punto di partenza del Capitolo VII.

Harada, T., Sasa, S. (2005). Equality connecting energy dissipation with a violation of the fluctuation-response relation. *Physical Review Letters*, 95(13), 130602. La relazione di Harada-Sasa per i sistemi fuori dall'equilibrio — la versione non-equilibrio del FDT citata nel paragrafo VII.5.

Relazioni di Onsager

Onsager, L. (1931). Reciprocal relations in irreversible processes. I. *Physical Review*, 37(4), 405-426; II. *Physical Review*, 38(12), 2265-2279. I due articoli originali delle relazioni di reciprocità — il risultato fondativo del Capitolo VIII.

de Groot, S.R., Mazur, P. (1962). *Non-Equilibrium Thermodynamics*. North-Holland, Amsterdam. Il testo classico della termodinamica lineare fuori dall'equilibrio — il framework del Capitolo VIII con le relazioni di Onsager e la loro applicazione ai sistemi di trasporto.

Teoria delle grandi deviazioni

Varadhan, S.R.S. (1966). Asymptotic probabilities and differential equations. *Communications on Pure and Applied Mathematics*, 19(3), 261-286. L'articolo fondativo della teoria delle grandi deviazioni di Varadhan.

Ellis, R.S. (1985). *Entropy, Large Deviations, and Statistical Mechanics*. Springer, New York. Il testo di riferimento per la teoria delle grandi deviazioni applicata alla meccanica statistica — la base del Capitolo IX.

Gärtner, J. (1977). On large deviations from the invariant measure. *Theory of Probability and Its Applications*, 22(1), 24-39. L'articolo originale del teorema di Gärtner, completato da Ellis (1984) per i processi di Markov.

Donsker, M.D., Varadhan, S.R.S. (1975). Asymptotic evaluation of certain Markov process expectations for large time. I–III. *Communications on Pure and Applied Mathematics*, 28-29. Il principio variazionale di Donsker-Varadhan per le grandi deviazioni dei processi di Markov — la base dell'Appendice G.

Instanton e meccanica semiclassica nelle grandi deviazioni

Freidlin, M.I., Wentzell, A.D. (1984). *Random Perturbations of Dynamical Systems*. Springer, New York. Il testo che sviluppa la teoria degli instanton per le grandi deviazioni dei sistemi dinamici perturbati — la base del calcolo dell'instanton cognitivo del paragrafo IX.5.

Vanden-Eijnden, E., Heymann, M. (2008). The geometric minimum action method: A least action principle on the space of curves. *Journal of Chemical Physics*, 128(6), 061103. Il metodo numerico per il calcolo degli instanton in sistemi di alta dimensione — lo strumento computazionale per trovare il percorso terapeutico ottimale del paragrafo IX.5.

Connessioni con la neuroscienza

Tishby, N., Pereira, F.C., Bialek, W. (1999). The information bottleneck method. In: *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 368-377. Il metodo dell'information bottleneck come compressione ottimale dell'informazione rilevante — connessione con la funzione di tasso per l'efficienza cognitiva.

Ganguli, S., Sompolinsky, H. (2012). Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annual Review of Neuroscience*, 35, 485-508. La compressione sparsa come meccanismo neurale di efficienza

dell'elaborazione dell'informazione — rilevante per il bound di Landauer cognitivo dell'Appendice B.